

# Data Provenance

---

David Koop  
Computer & Information Science  
University of Massachusetts Dartmouth

# Scientific Publication

## Galois Conjugates of Topological Phases

M. H. Freedman,<sup>1</sup> J. Gukelberger,<sup>2</sup> M. B. Hastings,<sup>1</sup> S. Trebst,<sup>1</sup> M. Troyer,<sup>2</sup> and Z. Wang<sup>1</sup>

<sup>1</sup>Microsoft Research, Station Q, University of California, Santa Barbara, CA 93106, USA

<sup>2</sup>Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland

(Dated: July 6, 2011)

Galois conjugation relates unitary conformal field theories (CFTs) and topological quantum field theories (TQFTs) to their non-unitary counterparts. Here we investigate Galois conjugates of quantum double models, such as the Levin-Wen model. While these Galois conjugated Hamiltonians are typically non-Hermitian, we find that their ground state wave functions still obey a generalized version of the usual code property (local operators do not act on the ground state manifold) and hence enjoy a generalized topological protection. The key question addressed in this paper is whether such non-unitary topological phases can also appear as the ground states of Hermitian Hamiltonians. Specific attempts at constructing Hermitian Hamiltonians with these ground states lead to a loss of the code property and topological protection of the degenerate ground states. Beyond this we rigorously prove that no local change of basis (IV.5) can transform the ground states of the Galois conjugated doubled Fibonacci theory into the ground states of a topological model whose Hermitian Hamiltonian satisfies Lieb-Robinson bounds. These include all gapped local or quasi-local Hamiltonians. A similar statement holds for many other non-unitary TQFTs. One consequence is that the “Gaffnian” wave function cannot be the ground state of a gapped fractional quantum Hall state.

PACS numbers: 05.30.Pr, 73.43.-f

### I. INTRODUCTION

*Galois conjugation*, by definition, replaces a root of a polynomial by another one with identical algebraic properties. For example,  $i$  and  $-i$  are Galois conjugate (consider  $z^2 + 1 = 0$ ) as are  $\phi = \frac{1+\sqrt{5}}{2}$  and  $-\frac{1}{\phi} = \frac{1-\sqrt{5}}{2}$  (consider  $z^2 - z - 1 = 0$ ), as well as  $\sqrt[3]{2}$ ,  $\sqrt[3]{2}e^{2\pi i/3}$ , and  $\sqrt[3]{2}e^{-2\pi i/3}$  (consider  $z^3 - 2 = 0$ ). In physics Galois conjugation can be used to convert non-unitary conformal field theories (CFTs) to unitary ones, and vice versa. One famous example is the non-unitary Yang-Lee CFT, which is Galois conjugate to the Fibonacci CFT  $(G_2)_1$ , the even (or integer-spin) subset of  $\text{su}(2)_3$ .

In statistical mechanics non-unitary conformal field theories have a venerable history.<sup>1,2</sup> However, it has remained less clear if there exist physical situations in which non-unitary models can provide a useful description of the low energy physics of a quantum mechanical system – after all, Galois conjugation typically destroys the Hermitian property of the Hamiltonian. Some non-Hermitian Hamiltonians, which surprisingly have totally real spectrum, have been found to arise in the study of  $PT$ -invariant one-particle systems<sup>3</sup> and in some Galois conjugate many-body systems<sup>4</sup> and might be seen to open the door a crack to the physical use of such models. Another situation, which has recently attracted some interest, is the question whether non-unitary models can describe 1D edge states of certain 2D bulk states (the edge holographic for the bulk). In particular, there is currently a discussion on whether or not the “Gaffnian” wave function could be the ground state for a *gapped* fractional quantum Hall (FQH) state albeit with a non-unitary “Yang-Lee” CFT describing its edge.<sup>5-7</sup> We conclude that this is not possible, further restricting the possible scope of non-unitary models in quantum mechanics.

We reach this conclusion quite indirectly. Our main thrust is the investigation of Galois conjugation in the simplest non-

Abelian Levin-Wen model.<sup>8</sup> This model, which is also called “DFib”, is a topological quantum field theory (TQFT) whose states are string-nets on a surface labeled by either a trivial or “Fibonacci” anyon. From this starting point, we give a rigorous argument that the “Gaffnian” ground state cannot be locally conjugated to the ground state of any topological phase, within a Hermitian model satisfying Lieb-Robinson (LR) bounds<sup>9</sup> (which includes but is not limited to gapped local and quasi-local Hamiltonians).

Lieb-Robinson bounds are a technical tool for local lattice models. In relativistically invariant field theories, the speed of light is a strict upper bound to the velocity of propagation. In lattice theories, the LR bounds provide a similar upper bound by a velocity called the LR velocity, but in contrast to the relativistic case there can be some exponentially small “leakage” outside the light-cone in the lattice case. The Lieb-Robinson bounds are a way of bounding the leakage outside the light-cone. The LR velocity is set by microscopic details of the Hamiltonian, such as the interaction strength and range. Combining the LR bounds with the spectral gap enables us to prove locality of various correlation and response functions. We will call a Hamiltonian a *Lieb-Robinson Hamiltonian* if it satisfies LR bounds.

We work primarily with a single example, but it should be clear that the concept of Galois conjugation can be widely applied to TQFTs. The essential idea is to retain the particle types and fusion rules of a unitary theory but when one comes to writing down the algebraic form of the  $F$ -matrices (also called  $6j$  symbols), the entries are now Galois conjugated. A slight complication, which is actually an asset, is that writing an  $F$ -matrix requires a gauge choice and the most convenient choice may differ before and after Galois conjugation.

Our method is not restricted to Galois conjugated DFib<sup>9</sup> and its factors Fib<sup>9</sup> and Fib<sup>9</sup>, but can be generalized to infinitely many non-unitary TQFTs, showing that they will not arise as low energy models for a gapped 2D quantum mechan-

### non-Hermitian DYL model

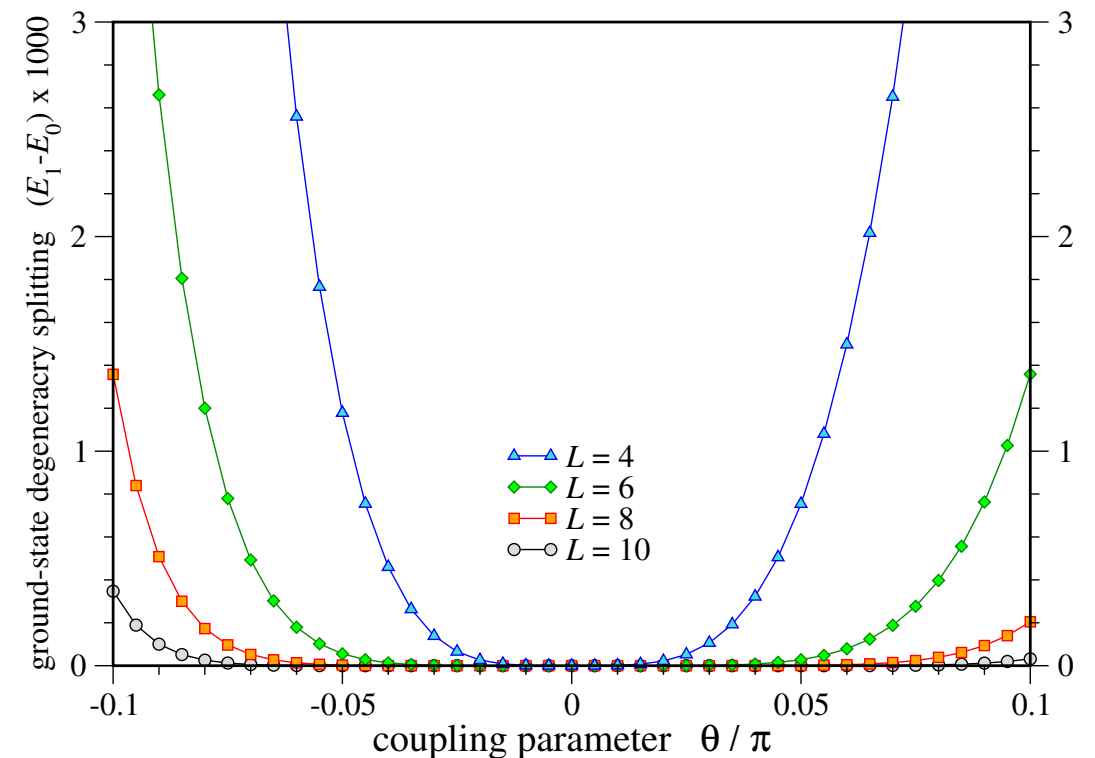
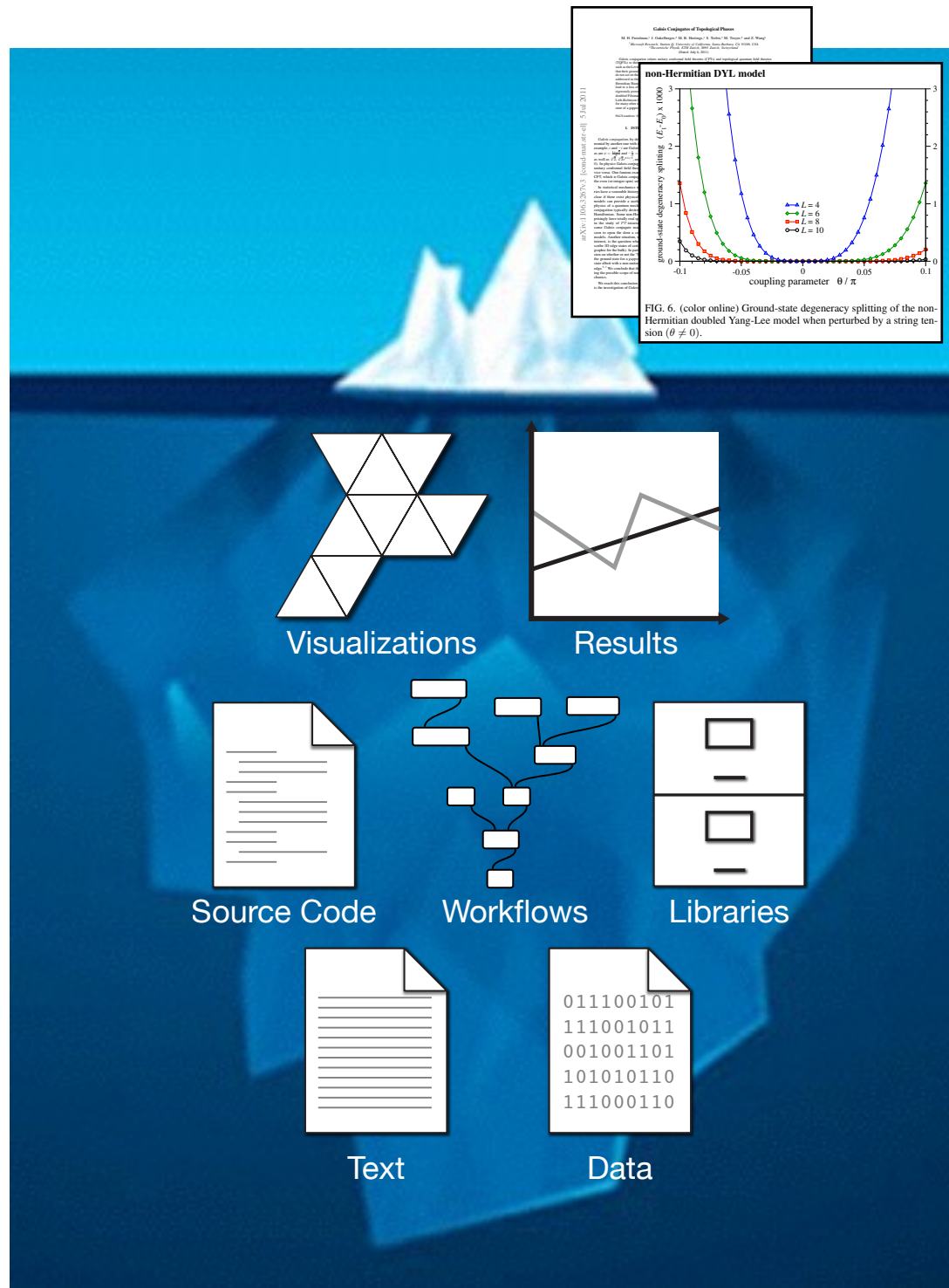


FIG. 6. (color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ( $\theta \neq 0$ ).

["Galois Conjugates of Topological Phases", Freedman et al., 2012]

# Scientific Publication



- Think about regenerating a figure from a paper written two years ago
  - Do you have the input data?
  - Do you know what software you used? Do you need to install it?
  - Does the software run on your current machine? Did the interfaces change?
  - Can you recreate the workflow/script?
  - How do you set the parameters?
  - Does everything run but you get a different result?
  - Is there a parameter/data file you forgot to record?



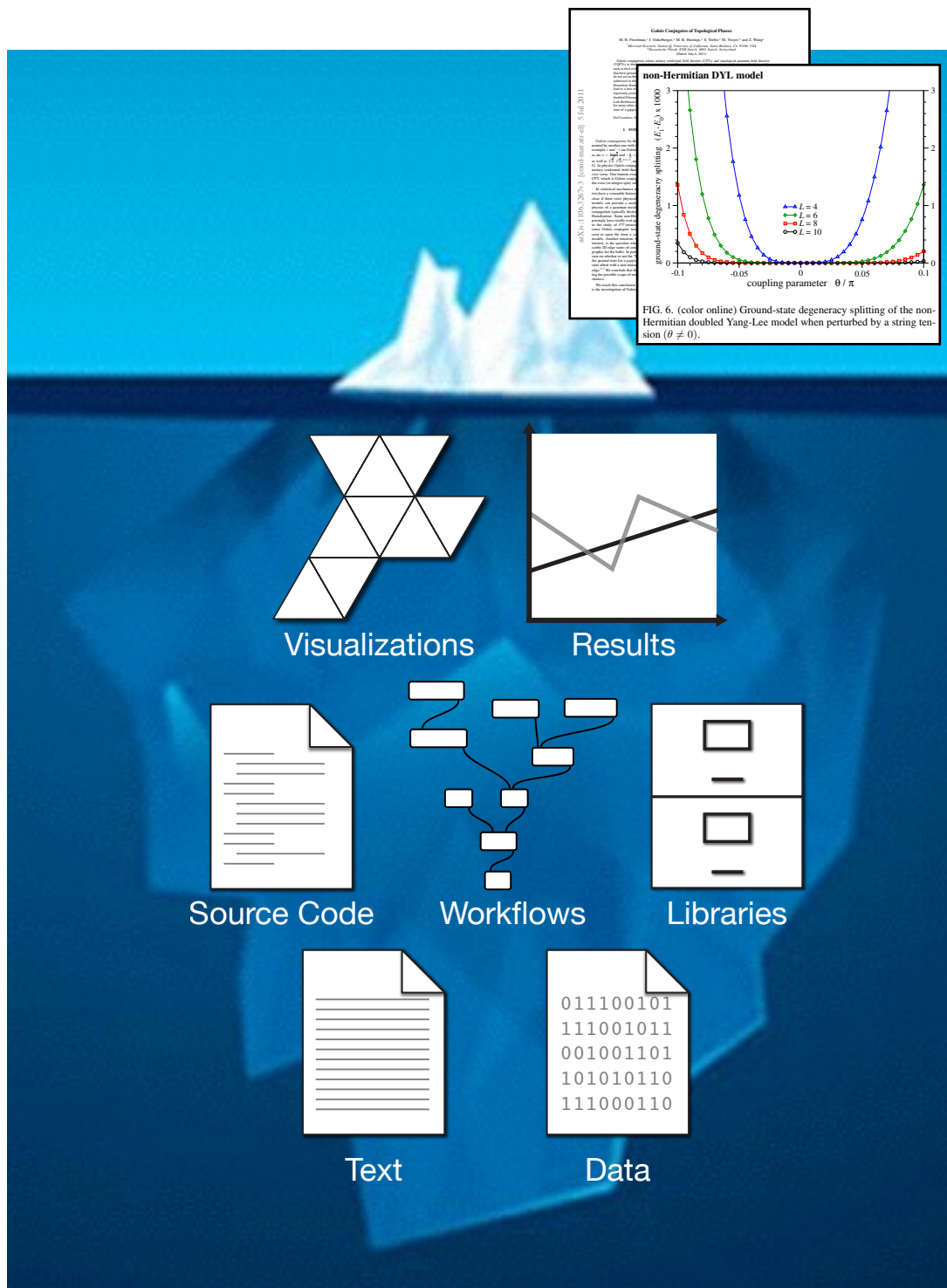
# Reproducibility



[Rube Goldberg Contest, Photo by Argonne National Laboratory, [CC BY-NC-SA 2.0](#)]



# Reproducibility



- Capture how results were achieved
- Includes many different items
- Improve **community** collaboration and sharing



# Ensure Quality, Reliability, & Trustworthiness

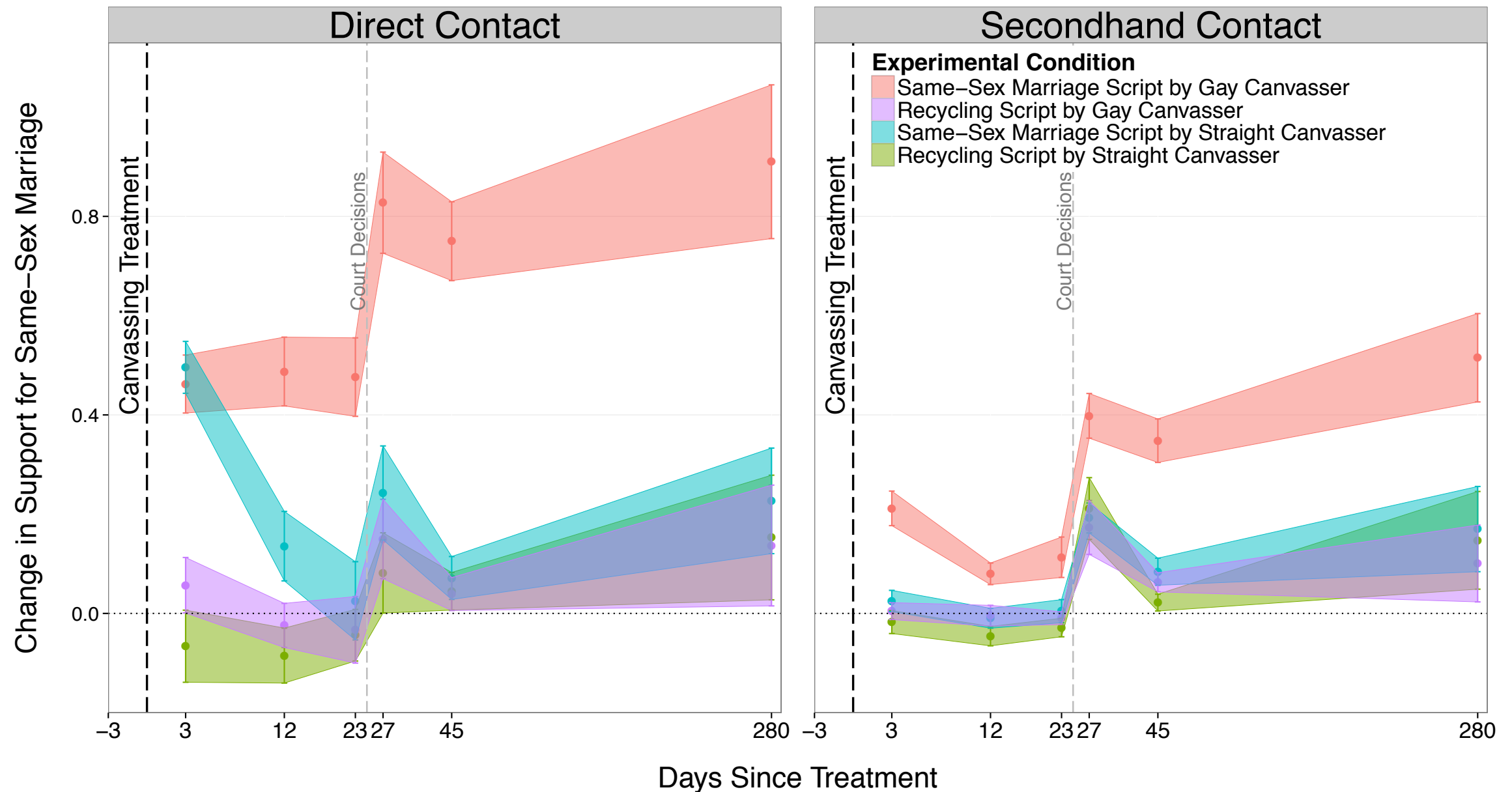
---



[Pentium Chip with FDIV Bug, Photo by Konstantin Lanzet, [CC BY-SA 3.0](#)]



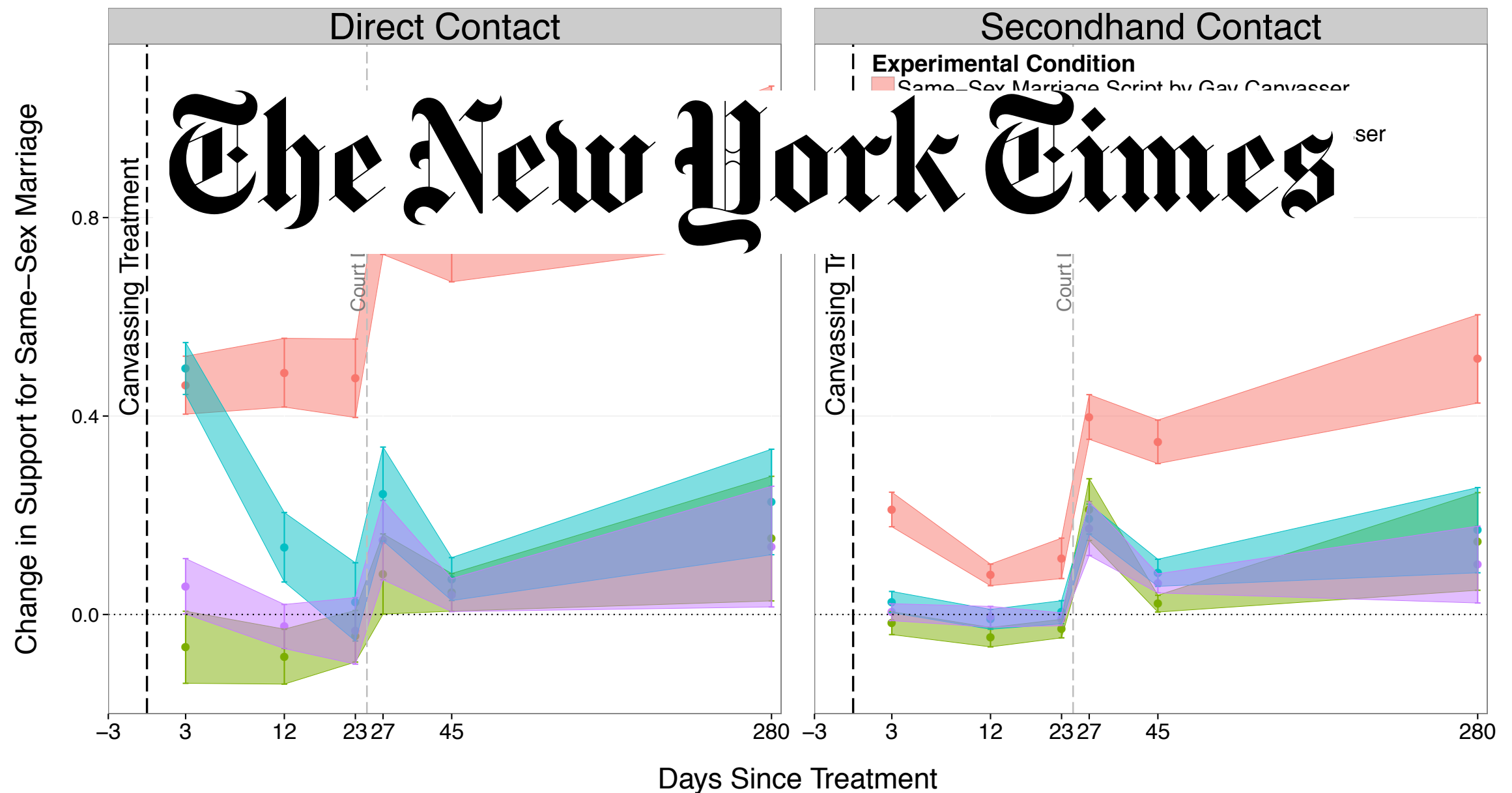
# Political Science Study in the News



*Note:* The first vertical dashed line represents the canvassing intervention, which was administered between Internet survey waves 1 and 2. The second vertical dashed line represents the U.S. Supreme Court decisions striking down California's ban on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves, with higher scores indicating more support for same-sex marriage. Points represent mean values, bars display 95% bootstrap confidence intervals.

[LaCour and Green, Science, 2014 (Retracted 2015)]

# Political Science Study in the News

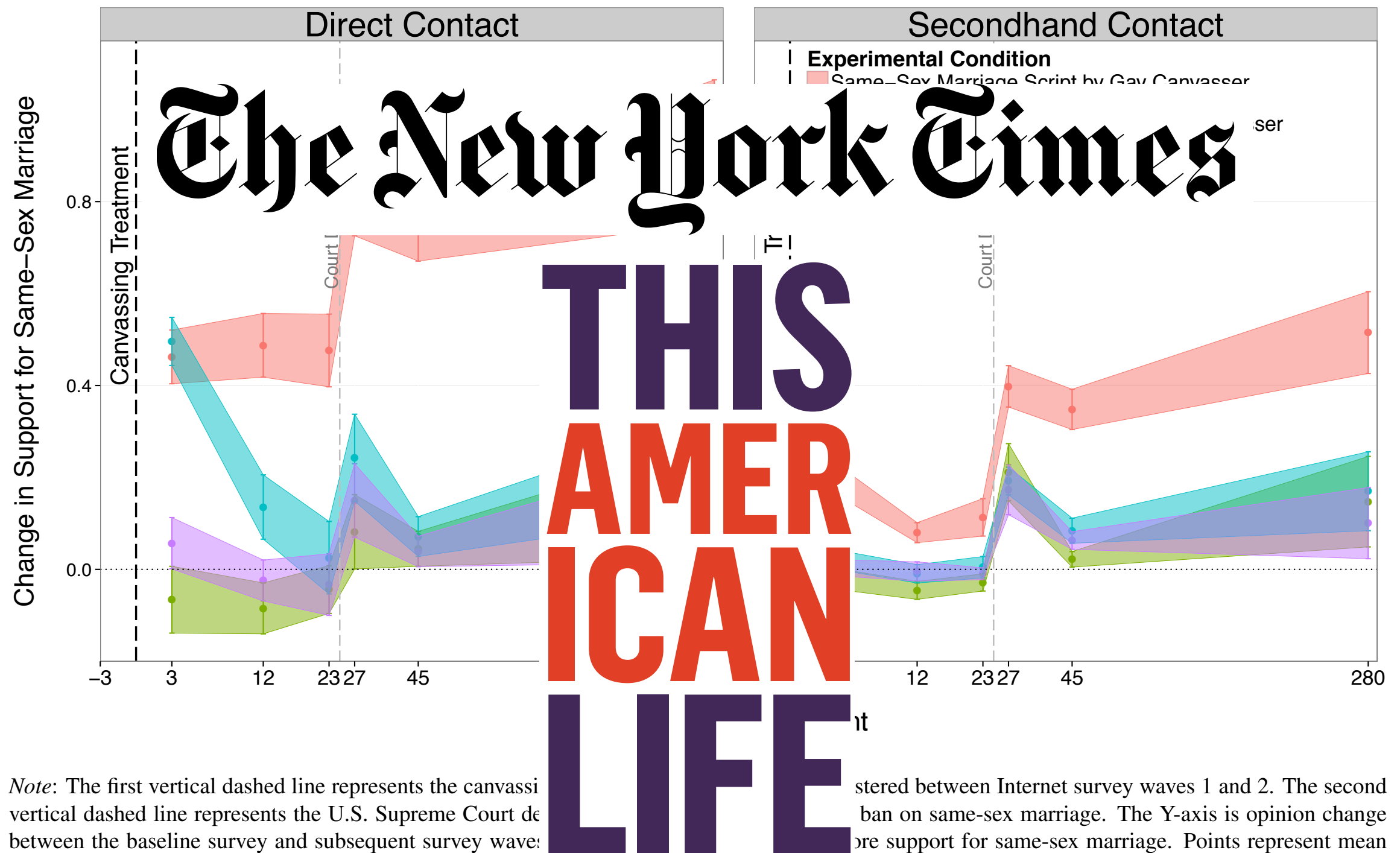


*Note:* The first vertical dashed line represents the canvassing intervention, which was administered between Internet survey waves 1 and 2. The second vertical dashed line represents the U.S. Supreme Court decisions striking down California's ban on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves, with higher scores indicating more support for same-sex marriage. Points represent mean values, bars display 95% bootstrap confidence intervals.

[LaCour and Green, Science, 2014 (Retracted 2015)]



# Political Science Study in the News



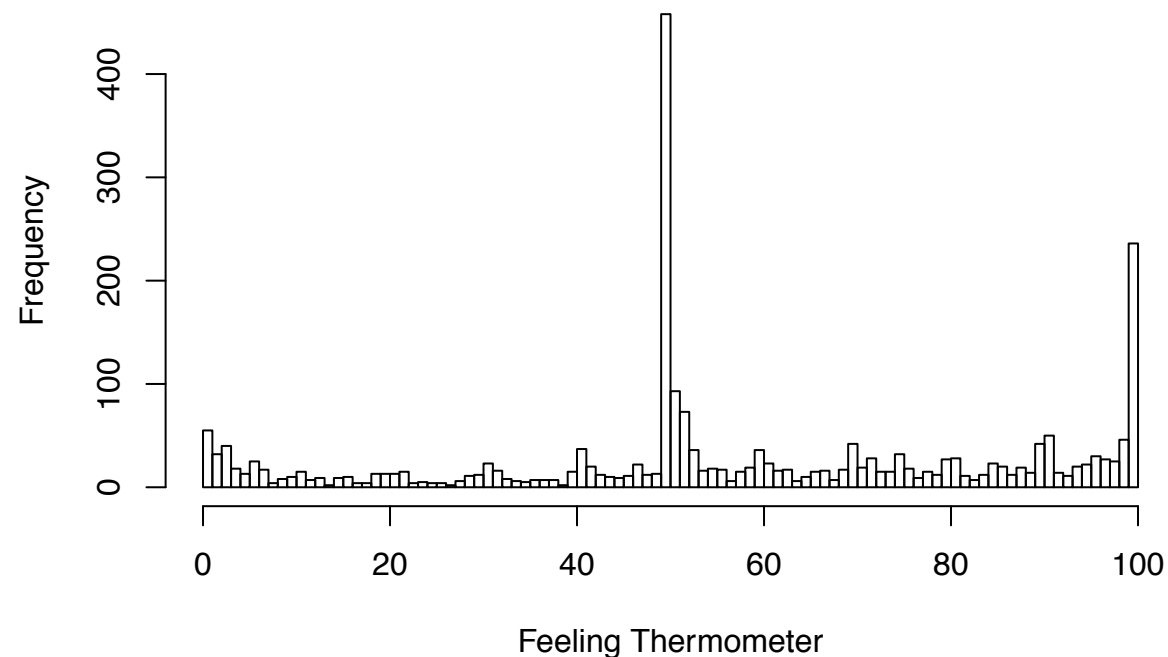
*Note:* The first vertical dashed line represents the canvassing treatment. The second vertical dashed line represents the U.S. Supreme Court decision on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves. Points represent mean values, bars display 95% bootstrap confidence intervals.

stered between Internet survey waves 1 and 2. The second ban on same-sex marriage. The Y-axis is opinion change between the baseline survey and subsequent survey waves. Points represent mean values, bars display 95% bootstrap confidence intervals.

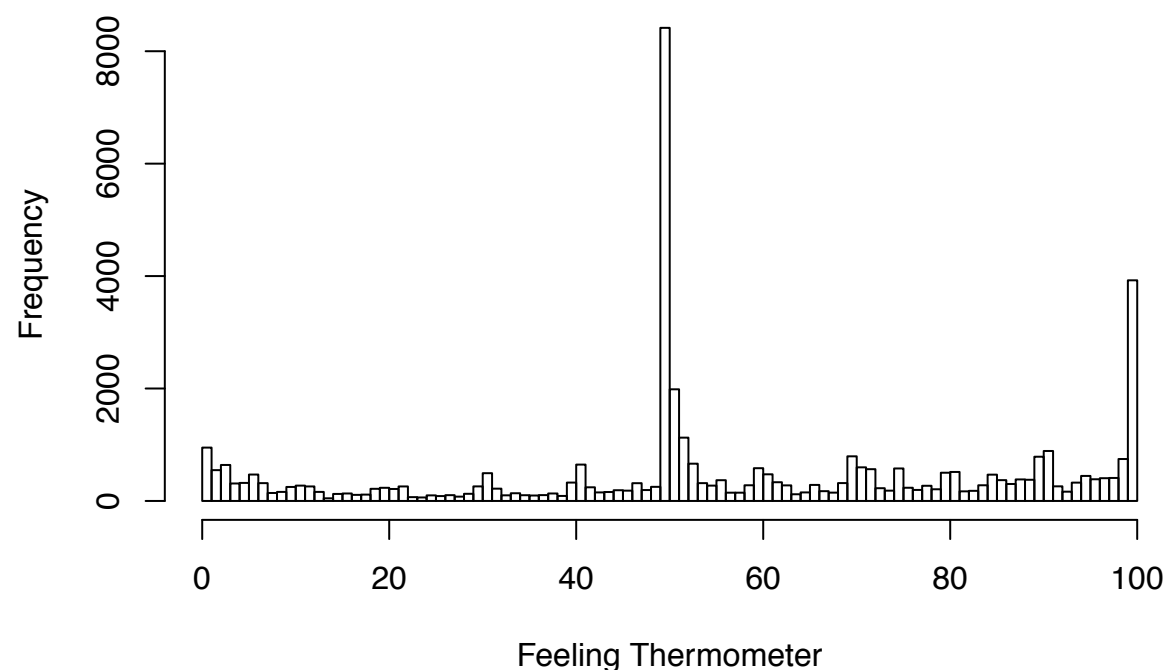
[LaCour and Green, Science, 2014 (Retracted 2015)]

# Reproducing Results

LaCour (2014) Study 2, Baseline



CCAP



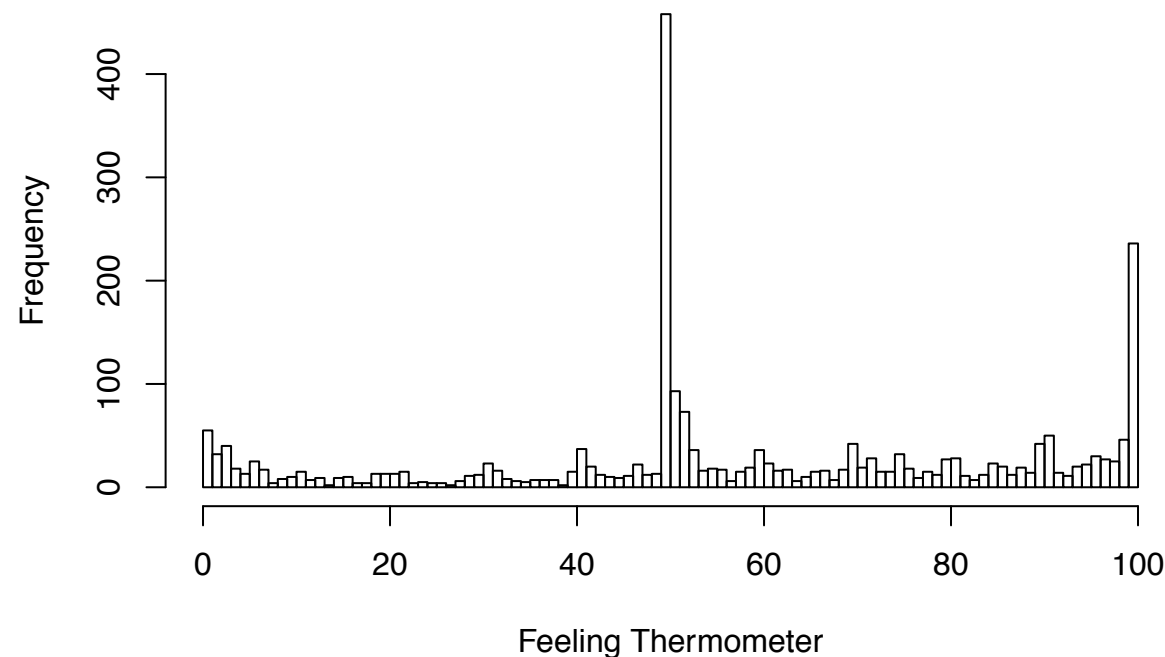
- "Irregularities in LaCour (2014)", Broockman, Kalla, and Aronow, 2015
- Tried their own pilot study and were unable to get similar results
- Found published data matched other data from different studies

[Broockman et al., 2015]



# Reproducing Results

LaCour (2014) Study 2, Baseline



- "Irregularities in LaCour (2014)", Broockman, Kalla, and Aronow, 2015
- Tried their own pilot study and were unable to get similar results
- Found published data matched other data from different studies

## Remaining Uncertainties

- We do not have access to the same-sex marriage question in CCAP, so we cannot evaluate the similarities of LaCour (2014)'s same-sex marriage question to the CCAP on that item.
- The claimed treatment effect was heterogeneous by canvasser attributes and the posted replication file does not have canvasser identifiers, so it is difficult to perform diagnostics on the responses of those assigned to treatment.
- The data for the abortion study reported at [http://www.cis.ethz.ch/content/dam/ethz/special-interest/gess/cis/cis-dam/CIS\\_DAM\\_2015/Colloquium/Papers/LaCour\\_2015.pdf](http://www.cis.ethz.ch/content/dam/ethz/special-interest/gess/cis/cis-dam/CIS_DAM_2015/Colloquium/Papers/LaCour_2015.pdf) in LaCour (2015) is not currently publicly available.

Feeling Thermometer

[Broockman et al., 2015]

# Retraction

Memo

May 19, 2015

To: Gilbert Chin

From: Donald Green

Re: Retraction of "LaCour, Michael J., and Donald P. Green. 2014. When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality. *Science*. 346(6215): 1366-1369."

I write to request a retraction of the above Science report. Last weekend, two UC Berkeley graduate students (David Broockman, and Josh Kalla) who had been working on a research project patterned after the studies reported in our article brought to my attention a series of irregularities that called into question the integrity of the data we present. They crafted a technical report with the assistance of Yale professor, Peter Aronow, and presented it to me last weekend. The report is attached. I brought their report to the attention of Lynn Vavreck, Professor of Political Science at UCLA and Michael LaCour's graduate advisor, who confronted him with these allegations on Monday morning, whereupon it was discovered that the on-line survey data that Michael LaCour purported to collect could not be traced to any originating Qualtrics source files. He claimed that he deleted the source file accidentally, but a Qualtrics service representative who examined the account and spoke with UCLA Political Science Department Chair Jeffrey Lewis reported to him that she found no evidence of such a deletion. On Tuesday, Professor Vavreck asked Michael LaCour for the contact information of survey respondents so that their participation in the survey could be verified, but he declined to furnish this information. With respect to the implementation of the surveys, Professor Vavreck was informed that, contrary to the description in the Supplemental Information, no cash incentives were offered or paid to respondents, and that, notwithstanding Michael LaCour's funding acknowledgement in the published report, he told Professor Vavreck that he did not in fact accept or use grant money to conduct surveys for either study, which she independently confirmed with the UCLA Law School and the UCLA Grants Office. Michael LaCour's failure to produce the raw data coupled with the other concerns noted above undermines the credibility of the findings.

I am deeply embarrassed by this turn of events and apologize to the editors, reviewers, and readers of Science.

- Green, one of the two authors, requested retraction after these questions arose
- "...survey data **could not be traced** to any originating Qualtrics source files" (emphasis added)
- "...**failure to produce the raw data** coupled with the other concerns noted above **undermines the credibility of the findings**" (emphasis added)
- Science retracted the paper on May 28, 2015



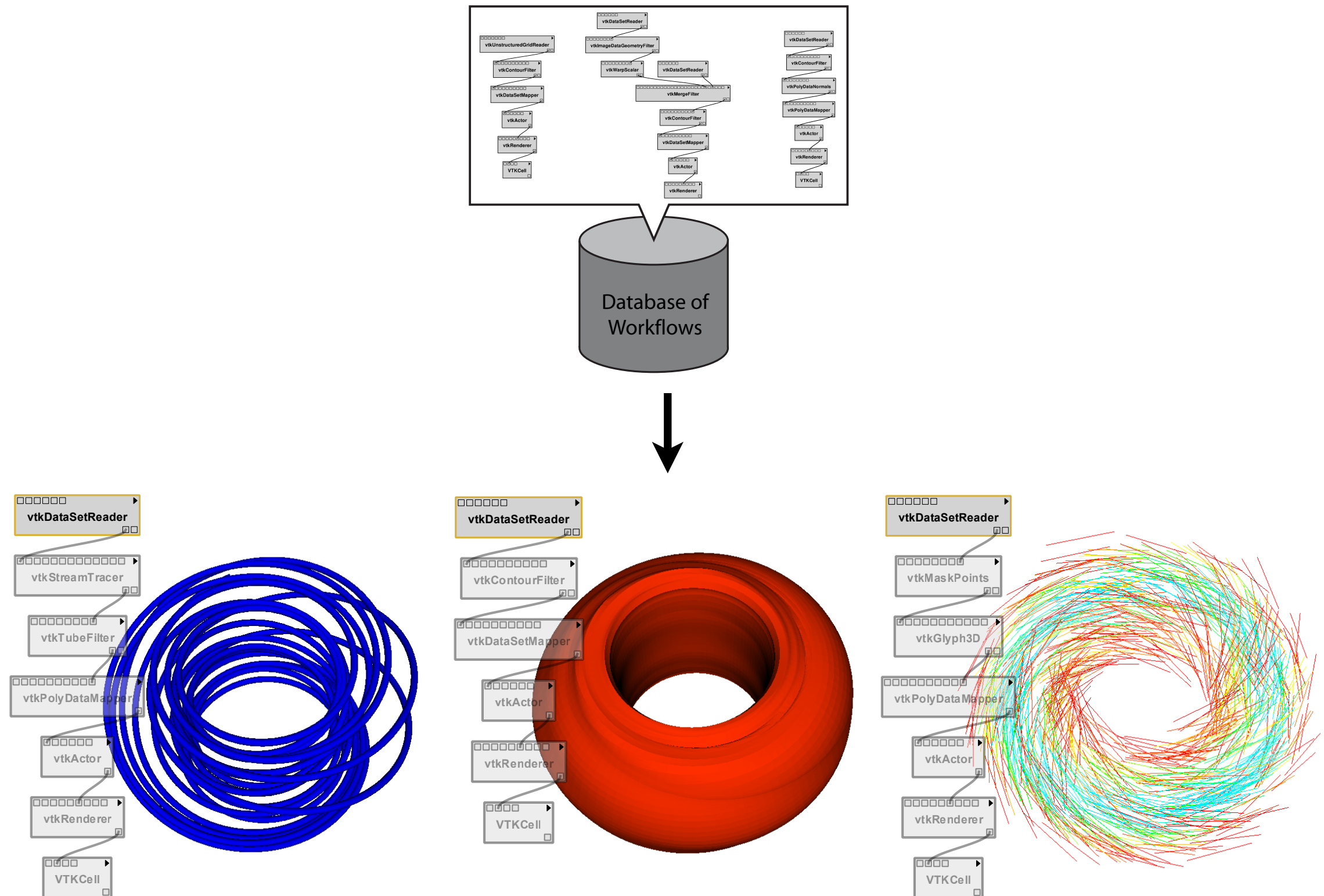
# Reuse Past Work



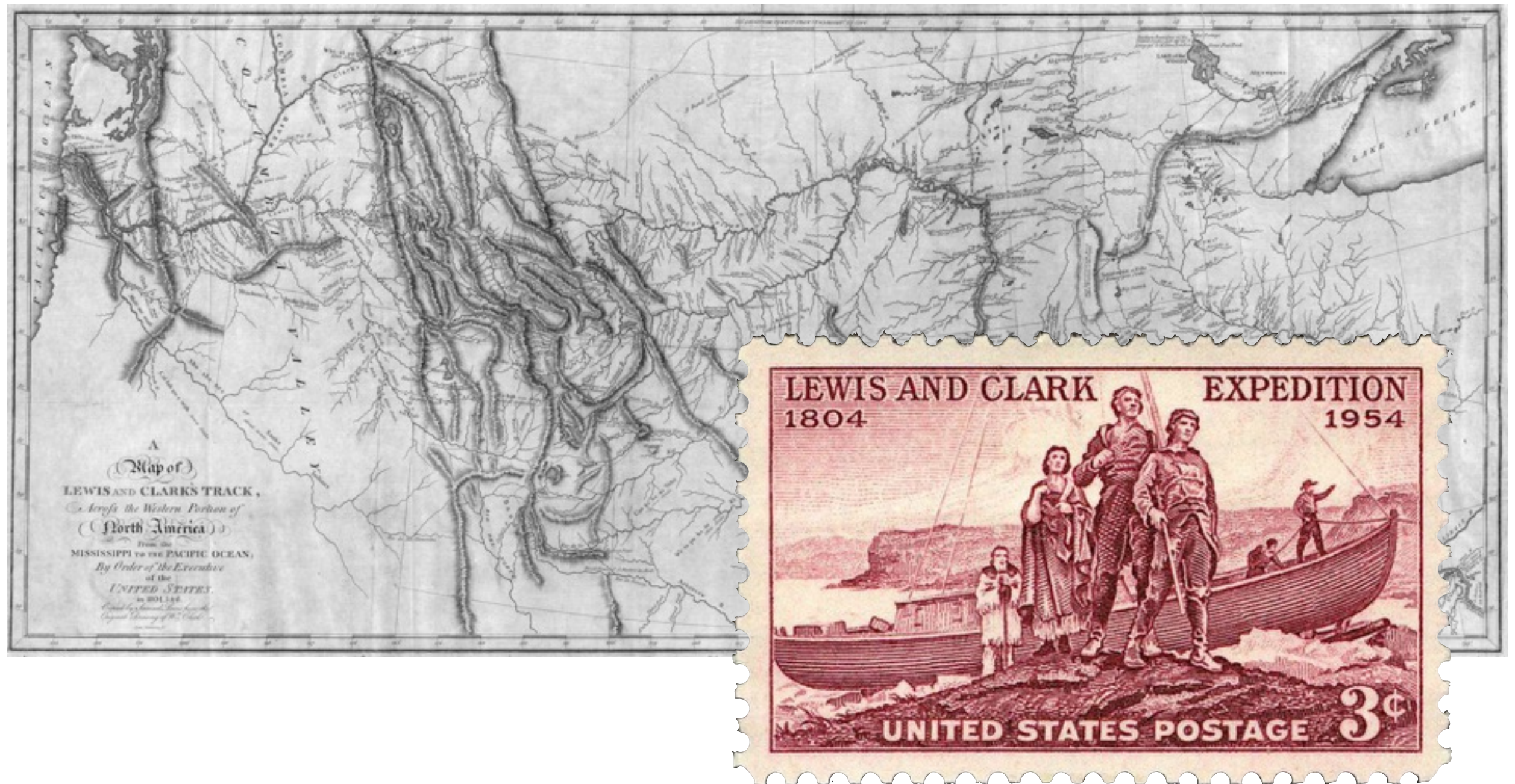
[PETCO Park, Photo by Edward O'Connor, CC BY-SA 2.0]



# Reuse Past Work



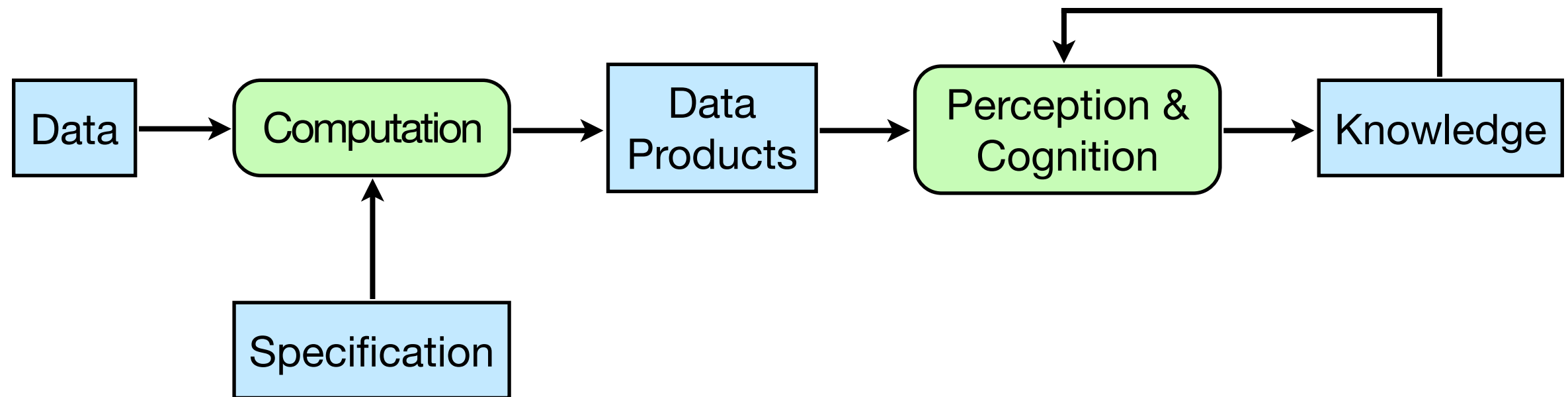
# Unencumbered Exploration & Learning





# Data Exploration

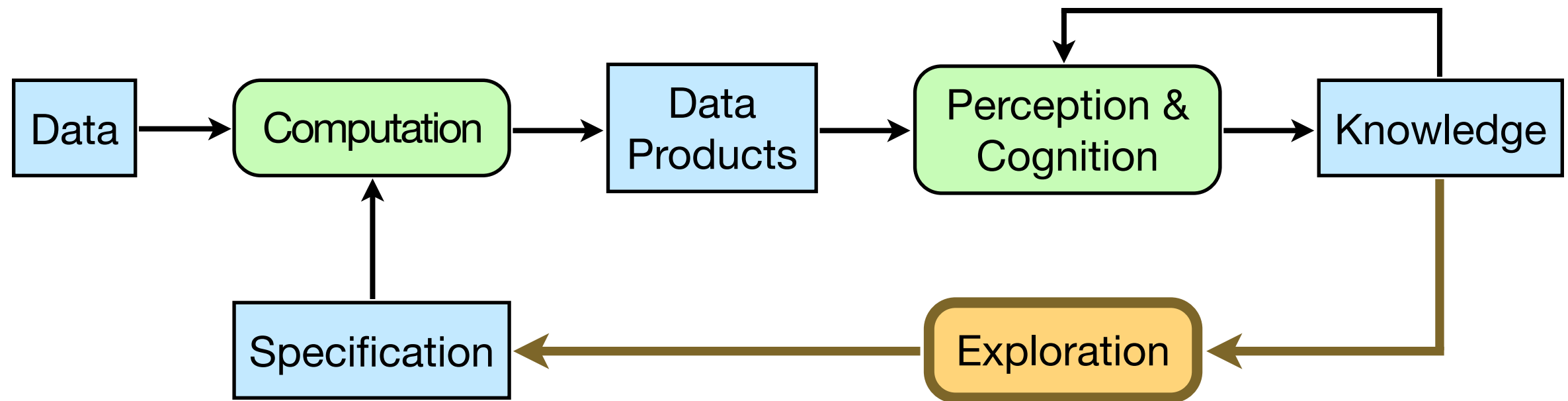
---



[Modified from Van Wijk, Vis 2005]

# Data Exploration

---



[Modified from Van Wijk, Vis 2005]

- Data analysis and visualization are iterative processes
- In exploratory tasks, change is the norm!

# Exploration and Creativity Support

---

- Reasoning is key to the exploratory processes
- "Reflective reasoning requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. ...the process is slow and laborious" — Donald A. Norman
- Need external aids—tools to facilitate this process
  - Creativity support tools [Ben Shneiderman]
- Need aid from people—collaboration



# Provenance in Art



## Rembrandt van Rijn

Dutch, 1606 - 1669

### ***Self-Portrait, 1659***

oil on canvas

Andrew W. Mellon Collection

1937.1.72

## Provenance

George, 3rd Duke of Montagu and 4th Earl of Cardigan [d. 1790], by 1767;[1] by inheritance to his daughter, Lady Elizabeth, wife of Henry, 3rd Duke of Buccleuch of Montagu House, London; John Charles, 7th Duke of Buccleuch; (P. & D. Colnaghi & Co., New York, 1928); (M. Knoedler & Co., New York); sold January 1929 to Andrew W. Mellon, Pittsburgh and Washington, D.C.; deeded 28 December 1934 to The A.W. Mellon Educational and Charitable Trust, Pittsburgh; gift 1937 to NGA.

[1] This early provenance is established by presence of a mezzotint after the portrait by R. Earlom (1743-1822), dated 1767. See John Charrington, *A Catalogue of the Mezzotints After, or Said to Be After, Rembrandt*, Cambridge, 1923, no. 49.

## Associated Names

- Buccleuch, Henry, 3rd Duke of
- Buccleuch, John Charles, 7th Duke of
- Colnaghi & Co., Ltd., P. & D.
- Knoedler & Company, M.
- Mellon, Andrew W.
- Mellon Educational and Charitable Trust, The A.W.
- Montagu, and 4th Earl of Cardigan, George, 3rd Duke of

[National Gallery of Art]

# Provenance in Art



## Rembrandt van Rijn

Dutch, 1606 - 1669

### ***Self-Portrait, 1659***

oil on canvas

Andrew W. Mellon Collection

1937.1.72

## Provenance

George, 3rd Duke of Montagu and 4th Earl of Cardigan [d. 1790], by 1767;[1] by inheritance to his daughter, Lady Elizabeth, wife of Henry, 3rd Duke of Buccleuch of Montagu House, London; John Charles, 7th Duke of Buccleuch; (P. & D. Colnaghi & Co., New York, 1928); (M. Knoedler & Co., New York); sold January 1929 to Andrew W. Mellon, Pittsburgh and Washington, D.C.; deeded 28 December 1934 to The A.W. Mellon Educational and Charitable Trust, Pittsburgh; gift 1937 to NGA.

[1] This early provenance is established by presence of a mezzotint after the portrait by R. Earlom (1743-1822), dated 1767. See John Charrington, *A Catalogue of the Mezzotints After, or Said to Be After, Rembrandt*, Cambridge, 1923, no. 49.

## Associated Names

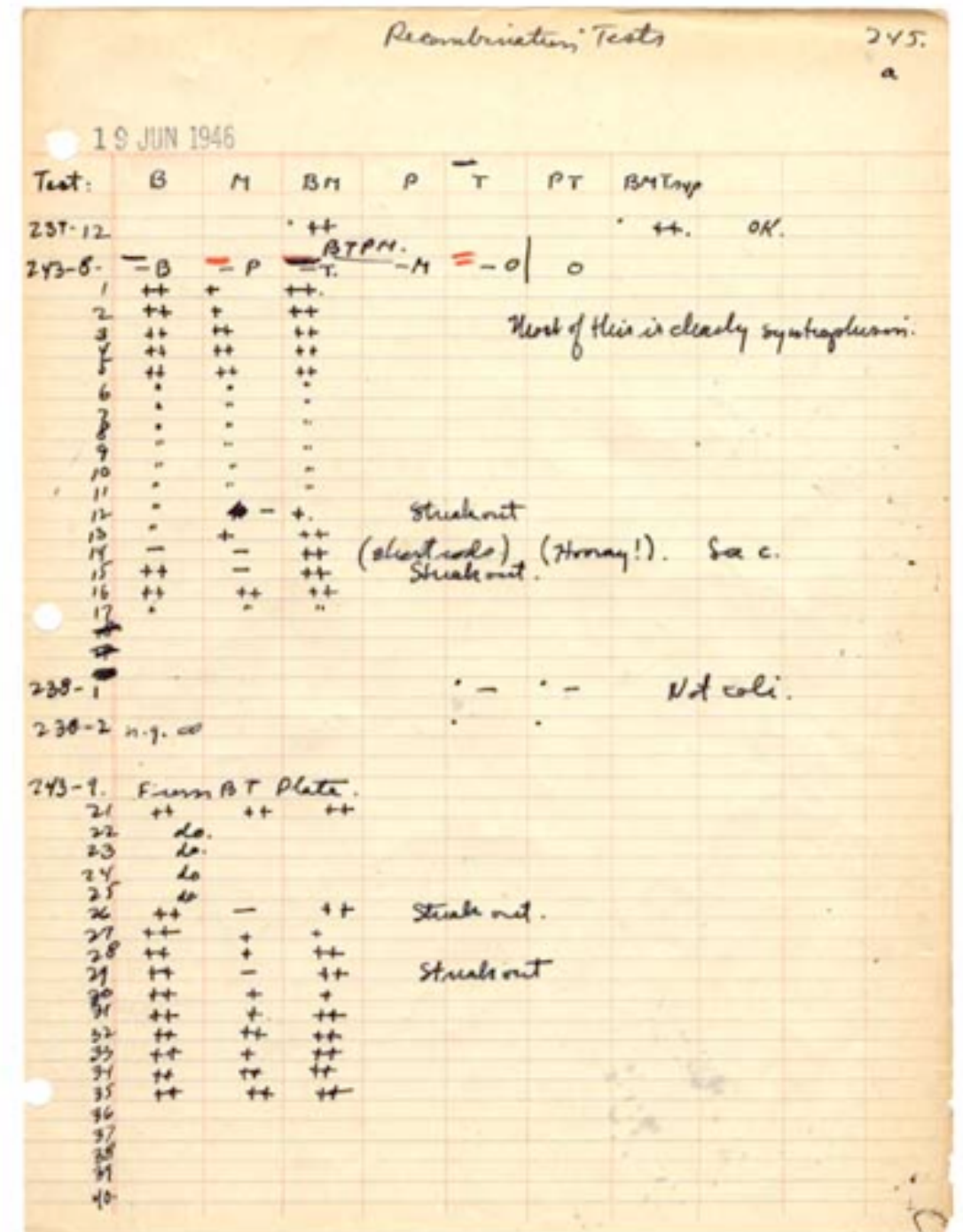
- Buccleuch, Henry, 3rd Duke of
- Buccleuch, John Charles, 7th Duke of
- Colnaghi & Co., Ltd., P. & D.
- Knoedler & Company, M.
- Mellon, Andrew W.
- Mellon Educational and Charitable Trust, The A.W.
- Montagu, and 4th Earl of Cardigan, George, 3rd Duke of

[National Gallery of Art]



# Provenance in Science

- Provenance: the lineage of data, a computation, or a visualization
- **Provenance is as (or more) important as the result!**
- Old solution:
  - Lab notebooks
- New problems:
  - Large volumes of data
  - Complex analyses
  - Writing notes doesn't scale

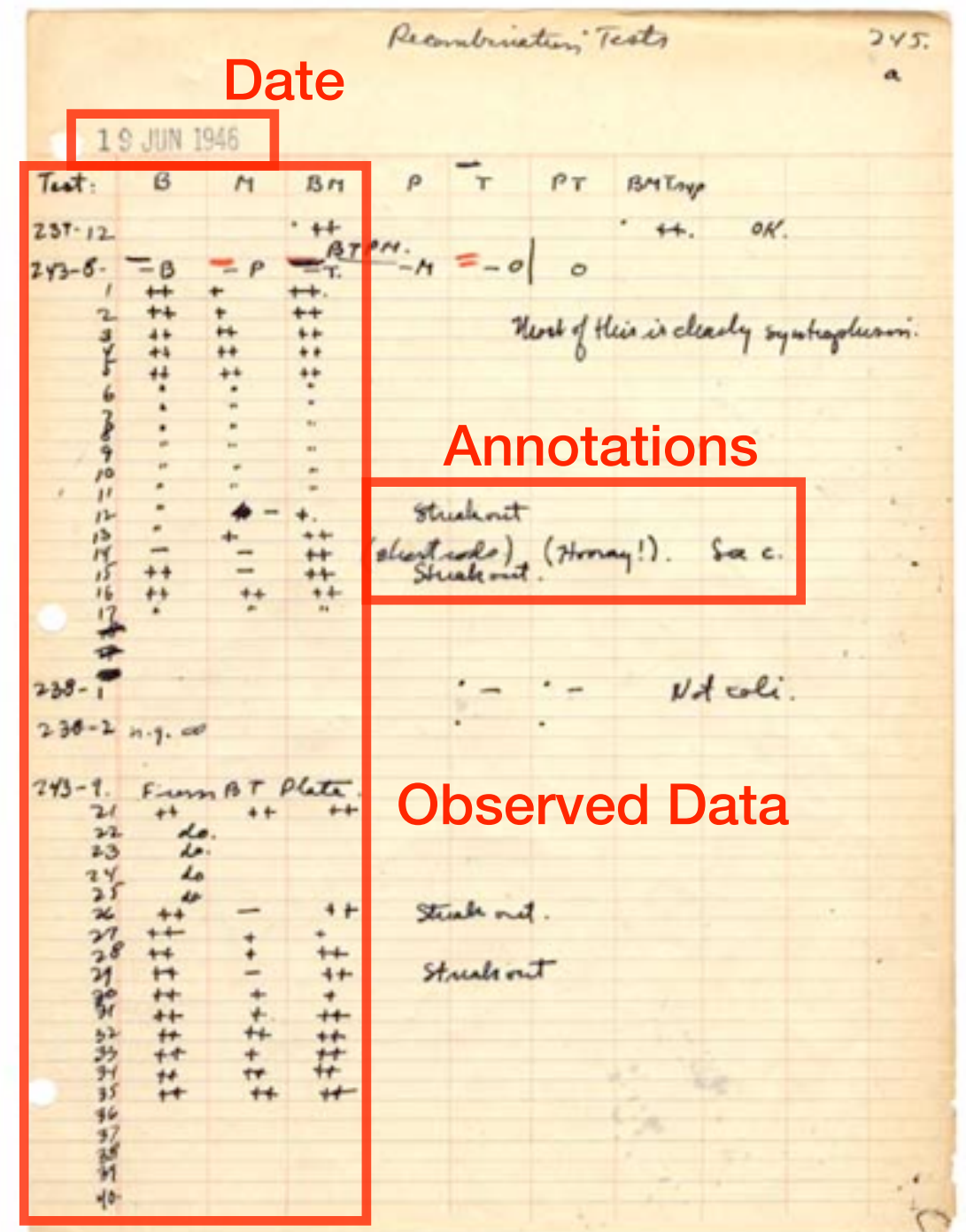


[DNA Recombination, Lederberg]



# Provenance in Science

- Provenance: the lineage of data, a computation, or a visualization
- **Provenance is as (or more) important as the result!**
- Old solution:
  - Lab notebooks
- New problems:
  - Large volumes of data
  - Complex analyses
  - Writing notes doesn't scale



[DNA Recombination, Lederberg]

# Provenance in Computational Science

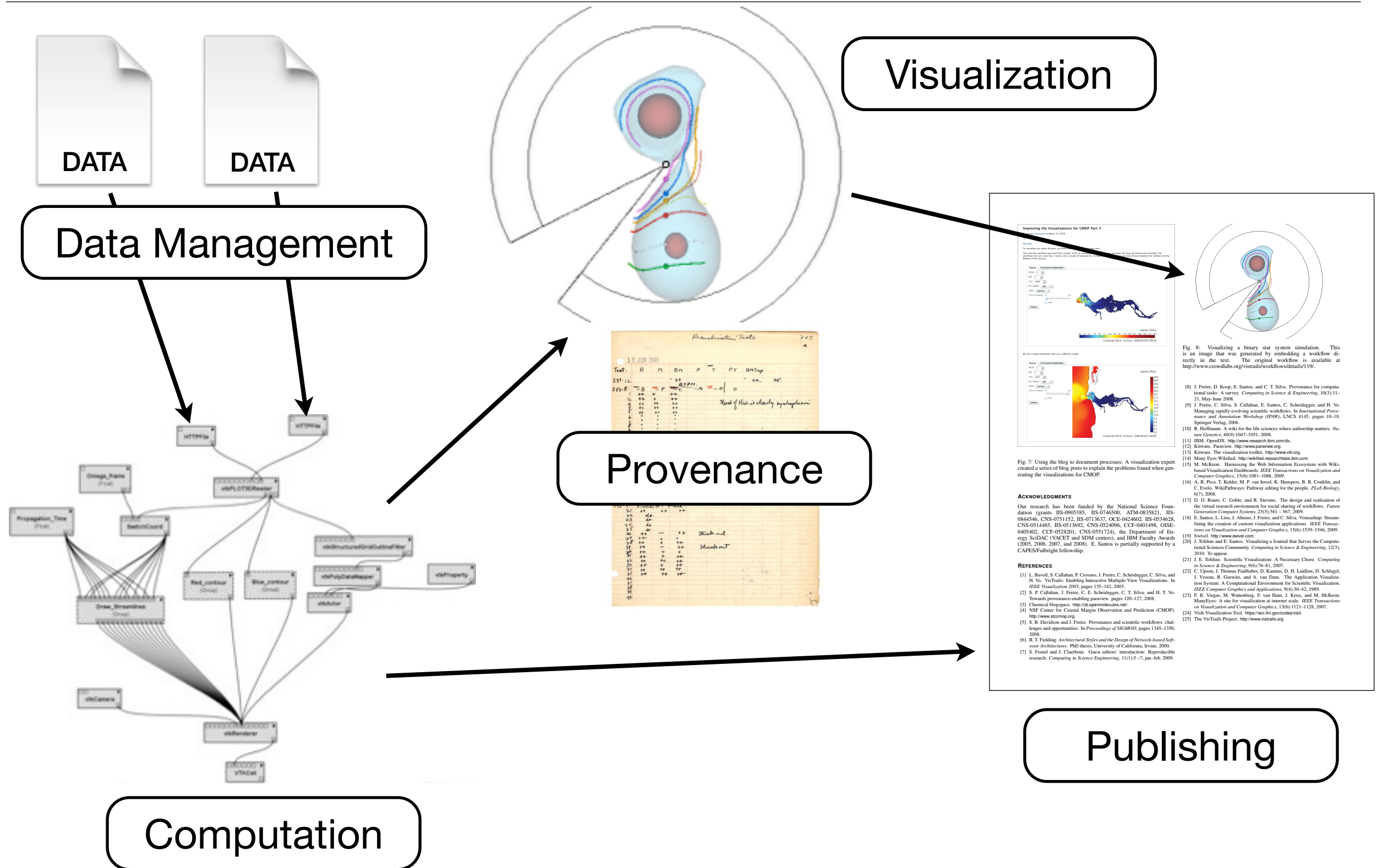


Fig. 8: Visualizing a binary star system simulation. This is an image that was generated by embedding a workflow directly in the text. The original workflow is available at <http://www.crowdmlabs.org/vitrails/workflows/details/119/>.

Fig. 7: Using the blog to document processes: A visualization expert created a series of blog posts to explain the problems found when generating the visualizations for CMOP.

**ACKNOWLEDGMENTS**

Our research has been funded by the National Science Foundation (grants IIS-0905385, IIS-0746500, ATM-0835821, IIS-0844546, CNS-0751152, IIS-0713637, CCE-0424602, IIS-0536628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, CCF-0528201, CNS-0551724), the Department of Energy SciDAC (VACET and SOM centers), and IBM Faculty Awards (2005, 2006, 2007, and 2008). E. Santos is partially supported by a CAPES/Fulbright fellowship.

**REFERENCES**

- [1] L. Barol, S. Callahan, P. Croso, J. Freire, C. Scheidegger, C. Silva, and H. Vo. Vitrails: Enabling Interactive Multiple-View Visualizations. In *IEEE Visualization 2005*, pages 135-142, 2005.
- [2] S. P. Callahan, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Towards provenance-enabling paraview. pages 120-127, 2008.
- [3] Chemical blogspace. <http://ch.opengemolecules.net/>.
- [4] NSF Center for Coastal Margin Observation and Prediction (CMOP). <http://www.stcmop.org>.
- [5] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of SIGMOD*, pages 1345-1350, 2008.
- [6] R. T. Fickling. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.
- [7] S. Finkel and J. Chelvar. Guest editors' introduction: Reproducible research. *Computing in Science Engineering*, 11(1):5-7, jan.-feb. 2009.
- [8] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11-21, May-June 2008.
- [9] J. Freire, C. Silva, S. Callahan, E. Santos, C. Scheidegger, and H. Vo. Managing rapidly-evolving scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10-18. Springer Verlag, 2008.
- [10] R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047-1051, 2008.
- [11] IBM. QsardX. <http://www.research.ibm.com/qsd/>.
- [12] Kitiware. Paraview. <http://www.paraview.org>.
- [13] Kitiware. The visualization toolkit. <http://www.vtk.org>.
- [14] Many Eyes Wikified. <http://wikified.researchgate.net.com>.
- [15] M. McKeon. Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1081-1088, 2009.
- [16] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hampers, B. R. Conklin, and C. Eickel. WikiPathways: Pathway editing for the people. *PLoS Biology*, 6(7), 2008.
- [17] D. D. Roure, C. Goble, and R. Stevens. The design and realization of the virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561-567, 2009.
- [18] E. Santos, L. Lima, J. Ahrens, J. Freire, and C. Silva. Vismashup: Streamlining the creation of custom visualization applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539-1546, 2009.
- [19] Svnedit. <http://www.svnedit.com>.
- [20] J. Tobline and E. Santos. Visualizing a Journal that Serves the Computational Sciences Community. *Computing in Science & Engineering*, 12(3), 2010. To appear.
- [21] J. E. Tobline. Scientific Visualization: A Necessary Chore. *Computing in Science & Engineering*, 9(6):76-81, 2007.
- [22] C. Upson, J. Thomas Fauthner, D. Kamin, D. H. Laidlaw, D. Schlegel, J. Vroom, R. Gurwitz, and A. van Dam. The Application Visualization System: A Computational Environment for Scientific Visualization. *IEEE Computer Graphics and Applications*, 9(4):30-42, 1989.
- [23] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121-1128, 2007.
- [24] Vitrails Visualization Tool. <http://wci.mit.gov/codes/vitrails/>.
- [25] The Vitrails Project. <http://www.vitrails.org>.

# Evolution of Publication

---

- Publish paper
- Publish code
- Publish computational experiments/tests
- Publish provenance (what actually happens during your runs)



# Provenance-Rich Publication

## Galois Conjugates of Topological Phases

M. H. Freedman,<sup>1</sup> J. Gukelberger,<sup>2</sup> M. B. Hastings,<sup>1</sup> S. Trebst,<sup>1</sup> M. Troyer,<sup>2</sup> and Z. Wang<sup>1</sup>

<sup>1</sup>Microsoft Research, Station Q, University of California, Santa Barbara, CA 93106, USA

<sup>2</sup>Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland

(Dated: July 6, 2011)

Galois conjugation relates unitary conformal field theories (CFTs) and topological quantum field theories (TQFTs) to their non-unitary counterparts. Here we investigate Galois conjugates of quantum double models, such as the Levin-Wen model. While these Galois conjugated Hamiltonians are typically non-Hermitian, we find that their ground state wave functions still obey a generalized version of the usual code property (local operators do not act on the ground state manifold) and hence enjoy a generalized topological protection. The key question addressed in this paper is whether such non-unitary topological phases can also appear as the ground states of Hermitian Hamiltonians. Specific attempts at constructing Hermitian Hamiltonians with these ground states lead to a loss of the code property and topological protection of the degenerate ground states. Beyond this we rigorously prove that no local change of basis (IV.5) can transform the ground states of the Galois conjugated doubled Fibonacci theory into the ground states of a topological model whose Hermitian Hamiltonian satisfies Lieb-Robinson bounds. These include all gapped local or quasi-local Hamiltonians. A similar statement holds for many other non-unitary TQFTs. One consequence is that the “Gaffnian” wave function cannot be the ground state of a gapped fractional quantum Hall state.

PACS numbers: 05.30.Pr, 73.43.-f

### I. INTRODUCTION

*Galois conjugation*, by definition, replaces a root of a polynomial by another one with identical algebraic properties. For example,  $i$  and  $-i$  are Galois conjugate (consider  $z^2 + 1 = 0$ ) as are  $\phi = \frac{1+\sqrt{5}}{2}$  and  $-\frac{1}{\phi} = \frac{1-\sqrt{5}}{2}$  (consider  $z^2 - z - 1 = 0$ ), as well as  $\sqrt[3]{2}$ ,  $\sqrt[3]{2}e^{2\pi i/3}$ , and  $\sqrt[3]{2}e^{-2\pi i/3}$  (consider  $z^3 - 2 = 0$ ). In physics Galois conjugation can be used to convert non-unitary conformal field theories (CFTs) to unitary ones, and vice versa. One famous example is the non-unitary Yang-Lee CFT, which is Galois conjugate to the Fibonacci CFT  $(G_2)_1$ , the even (or integer-spin) subset of  $su(2)_3$ .

In statistical mechanics non-unitary conformal field theories have a venerable history.<sup>1,2</sup> However, it has remained less clear if there exist physical situations in which non-unitary models can provide a useful description of the low energy physics of a quantum mechanical system – after all, Galois conjugation typically destroys the Hermitian property of the Hamiltonian. Some non-Hermitian Hamiltonians, which surprisingly have totally real spectrum, have been found to arise in the study of  $PT$ -invariant one-particle systems<sup>3</sup> and in some Galois conjugate many-body systems<sup>4</sup> and might be seen to open the door a crack to the physical use of such models. Another situation, which has recently attracted some interest, is the question whether non-unitary models can describe 1D edge states of certain 2D bulk states (the edge holographic for the bulk). In particular, there is currently a discussion on whether or not the “Gaffnian” wave function could be the ground state for a *gapped* fractional quantum Hall (FQH) state albeit with a non-unitary “Yang-Lee” CFT describing its edge.<sup>5-7</sup> We conclude that this is not possible, further restricting the possible scope of non-unitary models in quantum mechanics.

We reach this conclusion quite indirectly. Our main thrust is the investigation of Galois conjugation in the simplest non-

Abelian Levin-Wen model.<sup>8</sup> This model, which is also called “DFib”, is a topological quantum field theory (TQFT) whose states are string-nets on a surface labeled by either a trivial or “Fibonacci” anyon. From this starting point, we give a rigorous argument that the “Gaffnian” ground state cannot be locally conjugated to the ground state of any topological phase, within a Hermitian model satisfying Lieb-Robinson (LR) bounds<sup>9</sup> (which includes but is not limited to gapped local and quasi-local Hamiltonians).

Lieb-Robinson bounds are a technical tool for local lattice models. In relativistically invariant field theories, the speed of light is a strict upper bound to the velocity of propagation. In lattice theories, the LR bounds provide a similar upper bound by a velocity called the LR velocity, but in contrast to the relativistic case there can be some exponentially small “leakage” outside the light-cone in the lattice case. The Lieb-Robinson bounds are a way of bounding the leakage outside the light-cone. The LR velocity is set by microscopic details of the Hamiltonian, such as the interaction strength and range. Combining the LR bounds with the spectral gap enables us to prove locality of various correlation and response functions. We will call a Hamiltonian a *Lieb-Robinson Hamiltonian* if it satisfies LR bounds.

We work primarily with a single example, but it should be clear that the concept of Galois conjugation can be widely applied to TQFTs. The essential idea is to retain the particle types and fusion rules of a unitary theory but when one comes to writing down the algebraic form of the  $F$ -matrices (also called  $6j$  symbols), the entries are now Galois conjugated. A slight complication, which is actually an asset, is that writing an  $F$ -matrix requires a gauge choice and the most convenient choice may differ before and after Galois conjugation.

Our method is not restricted to Galois conjugated DFib<sup>9</sup> and its factors Fib<sup>9</sup> and Fib<sup>9</sup>, but can be generalized to infinitely many non-unitary TQFTs, showing that they will not arise as low energy models for a gapped 2D quantum mechan-

### non-Hermitian DYL model

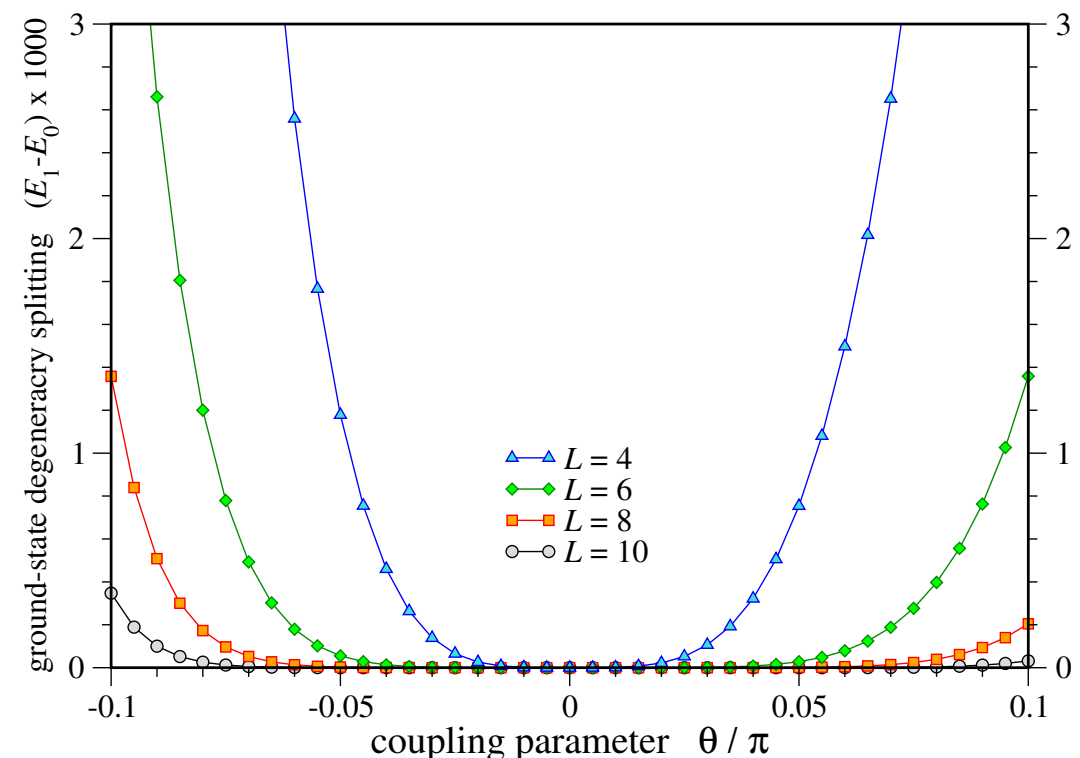


FIG. 6. (color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ( $\theta \neq 0$ ).

[Freedman et al., 2012]

# Benefits of Provenance-Rich Publications

---

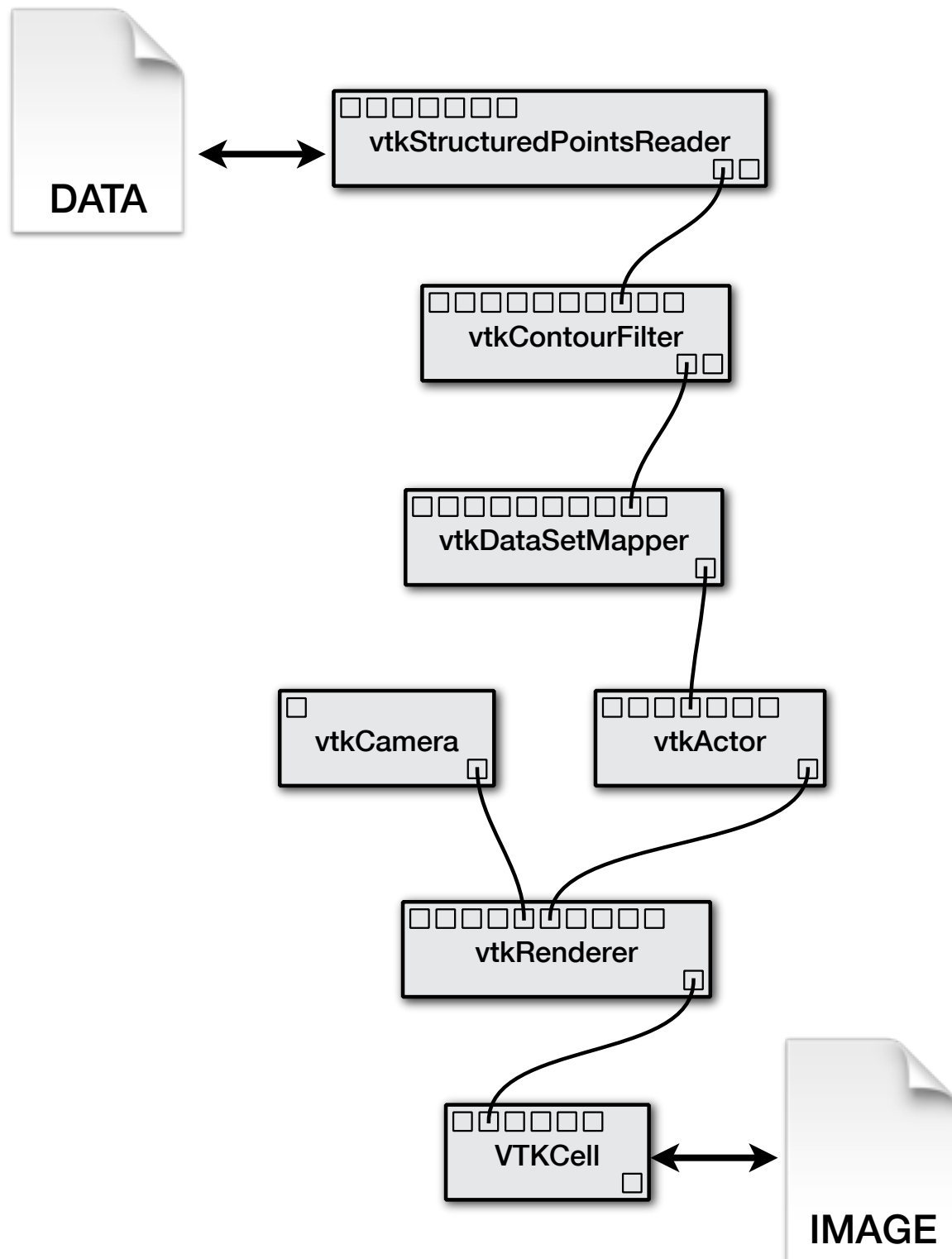
- Produce more knowledge—not just text
- Allow scientists to stand on the shoulders of giants (and their own)
- Science can move faster!
- Higher-quality publications
- Authors will be more careful
- Many eyes to check results
- Describe more of the discovery process: people only describe successes, can we learn from mistakes?
- Expose users to different techniques and tools: expedite their training; and potentially reduce their time to insight

# Provenance Definitions

---

- Dictionary: "the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners."
- Focus on **causality**—the sequence of steps that detail how a result was generated and/or **derivation**—what data a result depended on
- Provenance itself is **data**, this list of steps along with metadata for each step: when it occurred, who initiated it, notes about it
- Can be used to preserve information about an experiment and to answer many questions

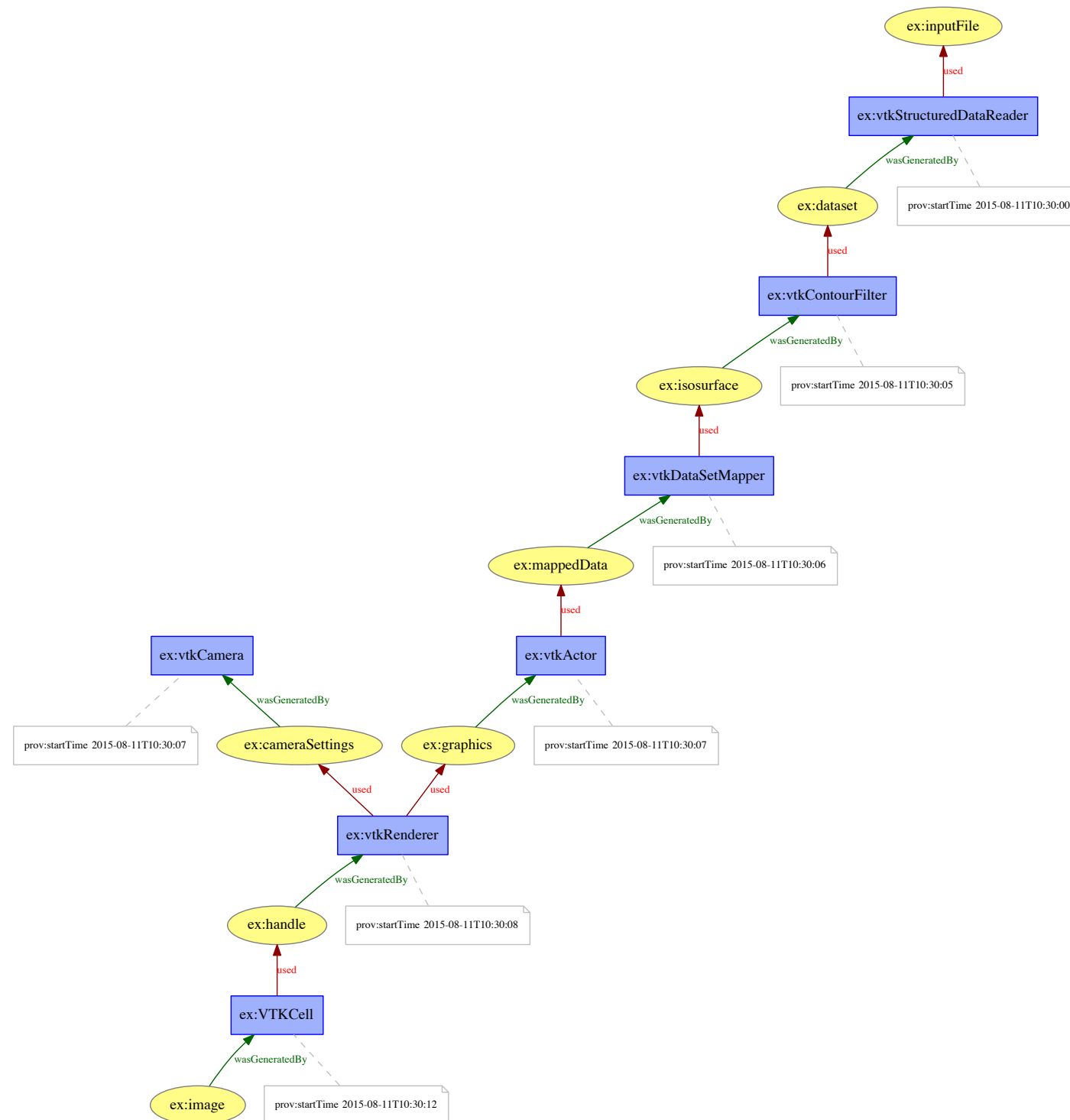
# Workflows



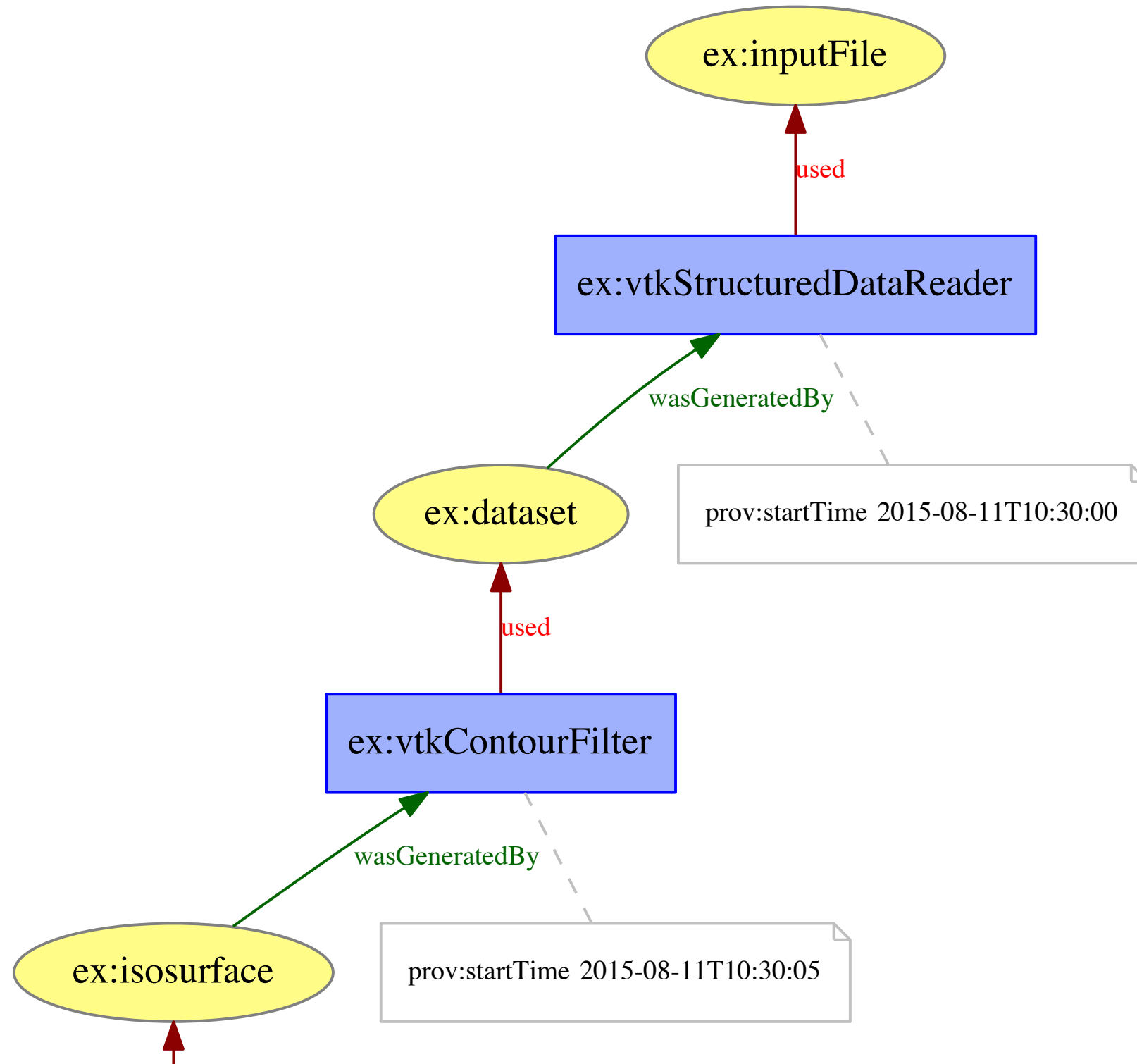
- Abstract computation
- Computational modules connected through input and output ports
- Data flows along the connections



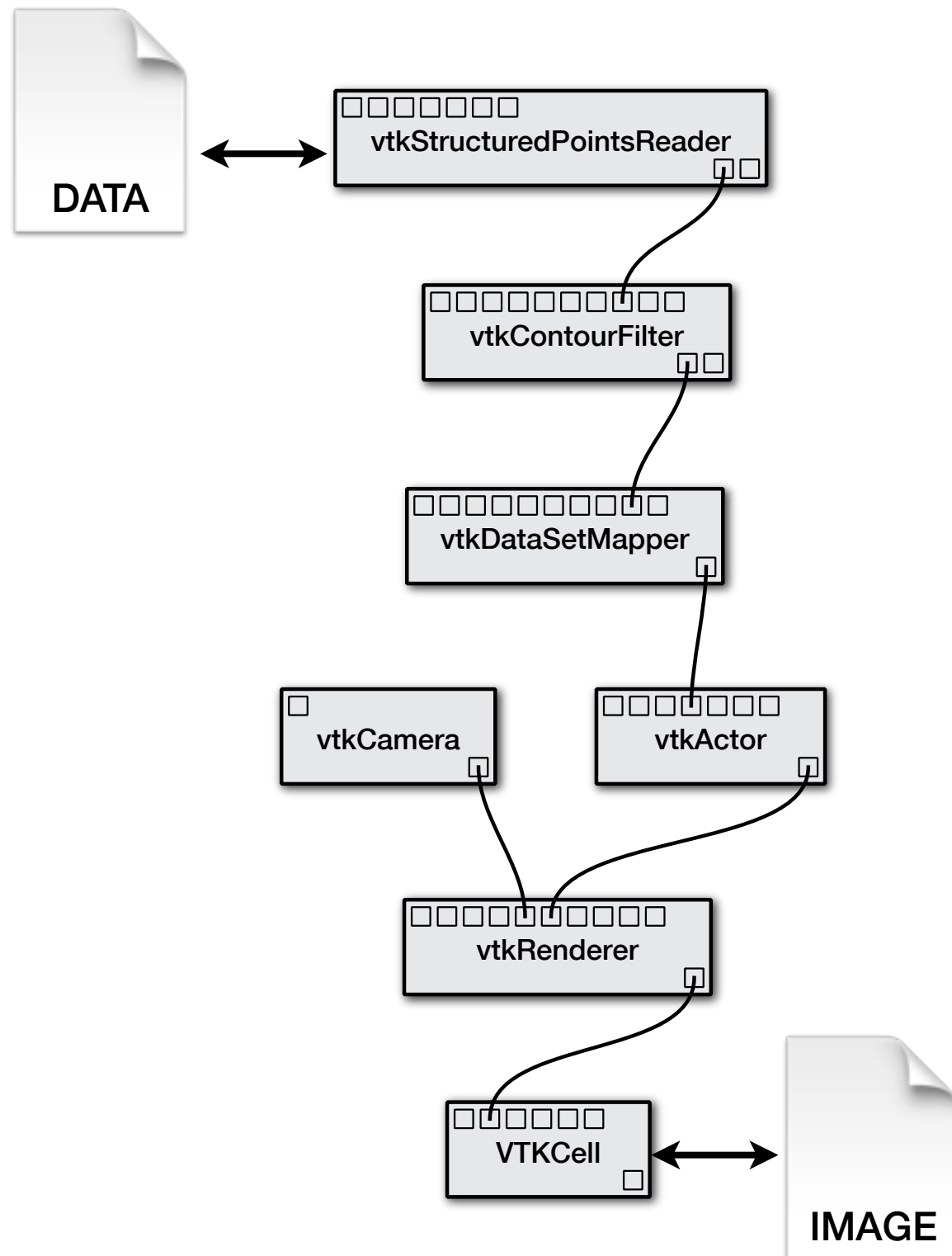
# Provenance Graph



# Provenance Graph



# Provenance Questions



- What process led to the output image?
- What input datasets contributed to the output image?
- What workflows create an isosurface with isovalue 57?
- Who create this data product?
- When was this data file created?
- Why was `vtkCamera` used?
- Why do two output images differ?

# Provenance Management

---

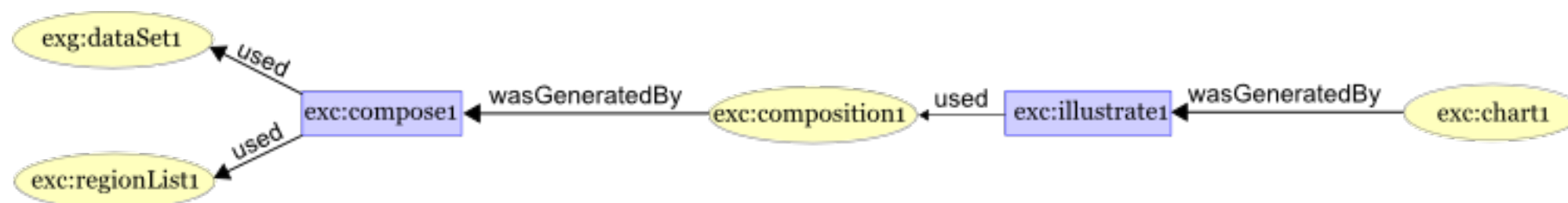
- Provenance can be generated from tasks/programs/scripts/etc.
- Properties of provenance is related to the **computational model**
  - a specific application with a graphical interface
  - a script that automates the use of several command-line tools
  - a scientific workflow that combines several tools



# Provenance & Causality

---

- Knowing what data/steps influenced other data/steps is important!
- Data dependencies: this output file depended on this input file
- Data-process dependencies: this output figure depended on these processes
- Causality can often be represented as a **graph** where connections represent dependencies



# User-defined provenance

---

- Goal: capture lots of provenance automatically based on what steps are executed
- Problem: not everything can be captured automatically
- Annotations offer ability to keep notes about processes
- Users might also specify known causal links that cannot be automatically determined (e.g. a step depends on three system files that were not specified as inputs in the workflow)

# Provenance Management

---

- What is needed to capture, store, and use provenance?
  1. Capture mechanism
  2. Model for representing provenance
  3. Tools to store, query, and analyze provenance

# Provenance Capture Mechanisms

---

- Workflow-based
  - Since workflow execution is controlled, keep track of all the workflow modules, parameters, etc. as they are executed
- Process-based
  - Each process is required to write out its own provenance information (not centralized like workflow-based)
- OS-based
  - The OS or filesystem is modified so that any activity it does it monitored and the provenance subsystem organizes it
- Tradeoffs:
  - Workflow- and process-based have better abstraction, OS-based requires minimal user effort once installed and can capture "hidden dependencies"



# Provenance Granularity

---

- How detailed should our provenance be?
  - Coarse: "This program ran with inputs x, y, z and produced outputs a, b, c"
  - Fine: "Input x was read into register 4, input y was read in register 5, add operation was performed using registers 4 and 5, ..."
- More queries are possible with fine-grained provenance, but...
  - Storage concerns
  - Performance concerns
- Abstraction can help here

# Abstraction: Script, Workflow, Abstract Workflow

```
data = vtk.vtkStructuredPointsReader()
data.SetFileName("../examples/data/head.120.vtk")

contour = vtk.vtkContourFilter()
contour.SetInput(data.GetOutput())
contour.SetValue(0, 67)

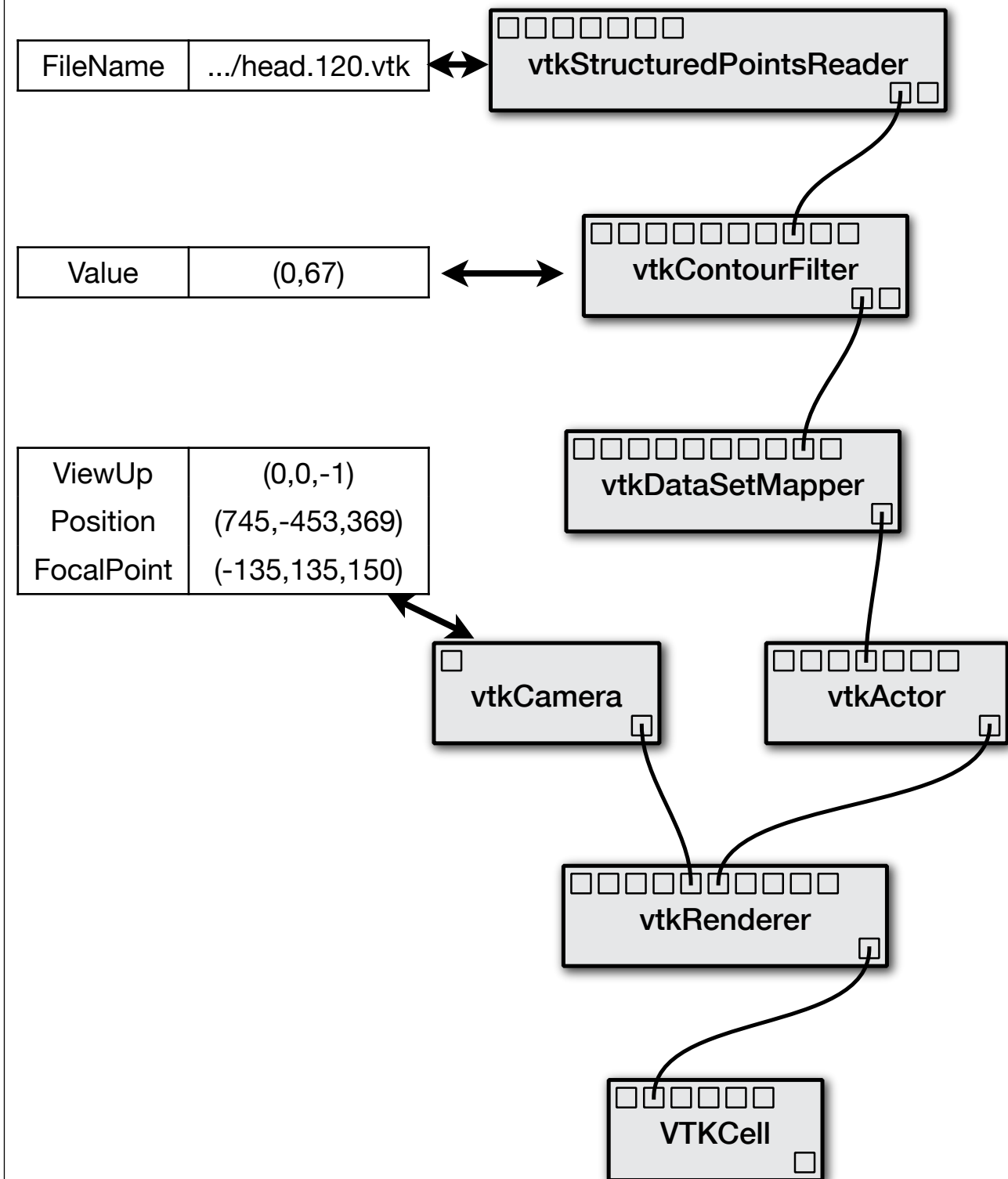
mapper = vtk.vtkPolyDataMapper()
mapper.SetInput(contour.GetOutput())
mapper.ScalarVisibilityOff()

actor = vtk.vtkActor()
actor.SetMapper(mapper)

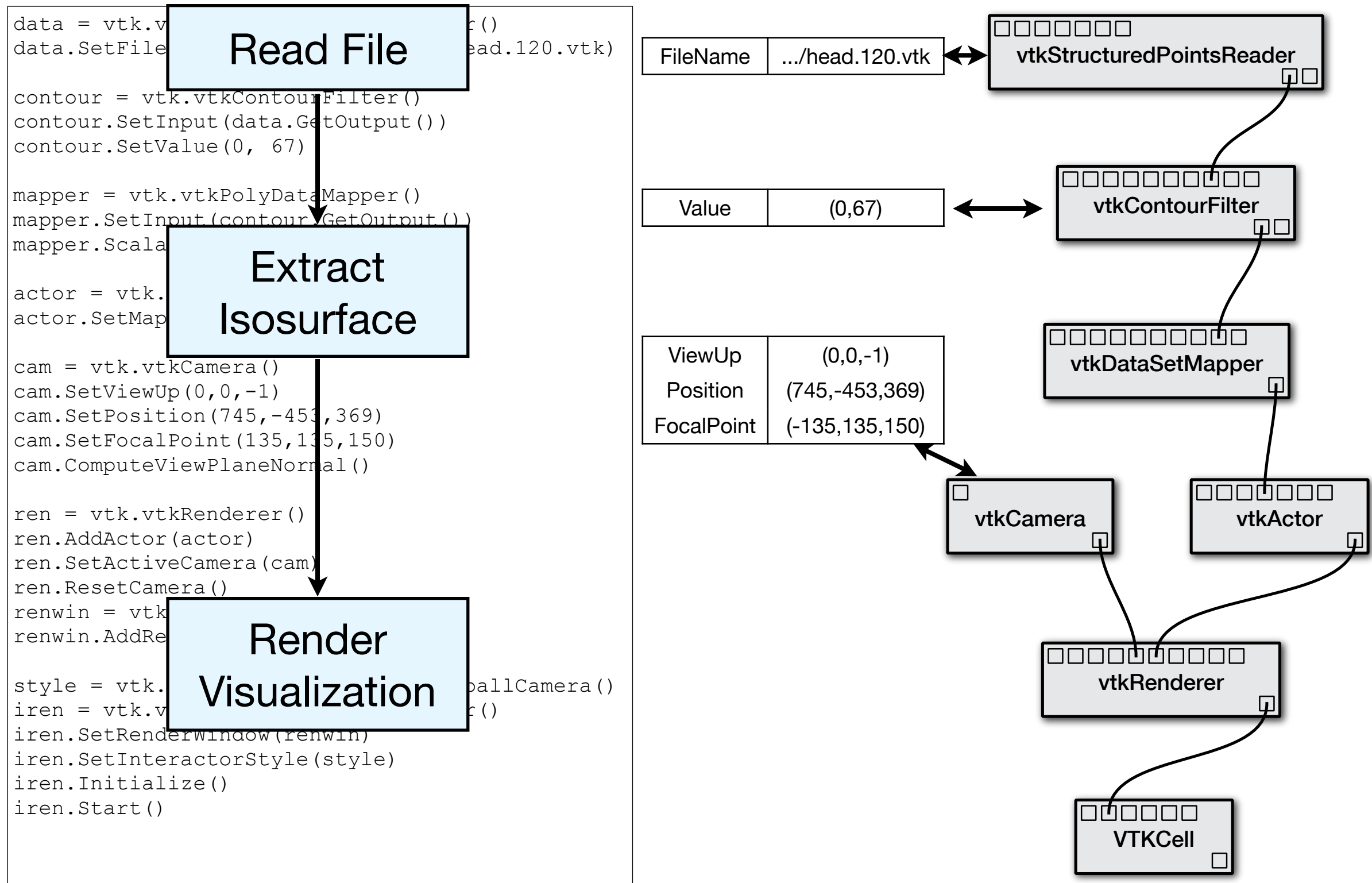
cam = vtk.vtkCamera()
cam.SetViewUp(0, 0, -1)
cam.SetPosition(745, -453, 369)
cam.SetFocalPoint(135, 135, 150)
cam.ComputeViewPlaneNormal()

ren = vtk.vtkRenderer()
ren.AddActor(actor)
ren.SetActiveCamera(cam)
ren.ResetCamera()
renwin = vtk.vtkRenderWindow()
renwin.AddRenderer(ren)

style = vtk.vtkInteractorStyleTrackballCamera()
iren = vtk.vtkRenderWindowInteractor()
iren.SetRenderWindow(renwin)
iren.SetInteractorStyle(style)
iren.Initialize()
iren.Start()
```

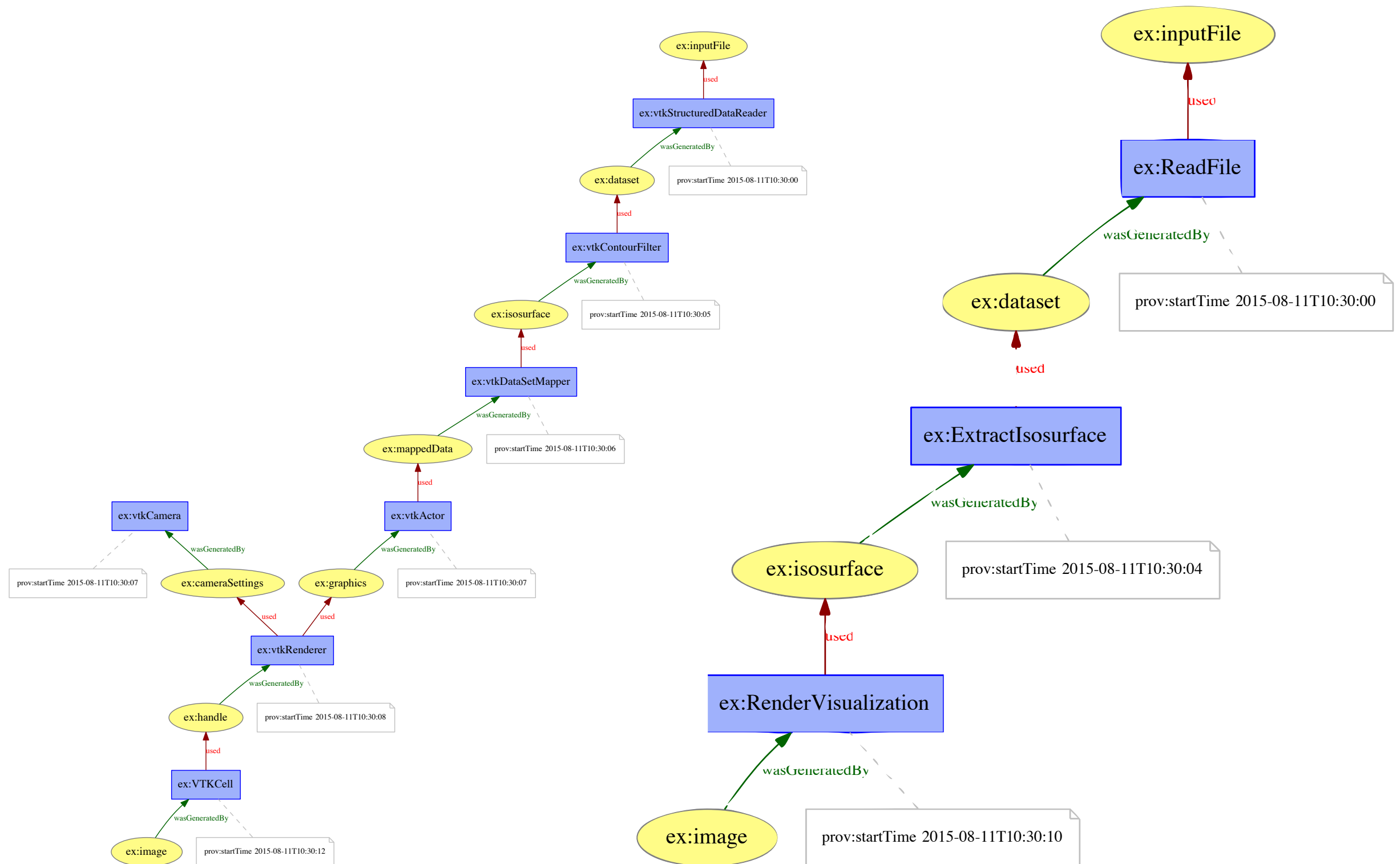


# Abstraction: Script, Workflow, Abstract Workflow





# Abstraction: Provenance Views



# Provenance Storage

---

- Keeping provenance for each data item means lots of **repetition**
- Nested data storage also induces repetition
- Coarse provenance is naturally more compact, but how to decide what (not) to store?
- Repeated provenance is not uncommon:
  - Repeating the same computation with a different parameter
  - Creating a new computation that has a very similar structure to one that was run two weeks ago
- Provenance compression/factorization techniques (e.g. [Chapman et al., 2008], [Anand et al., 2009]) take advantage of that to reduce storage costs

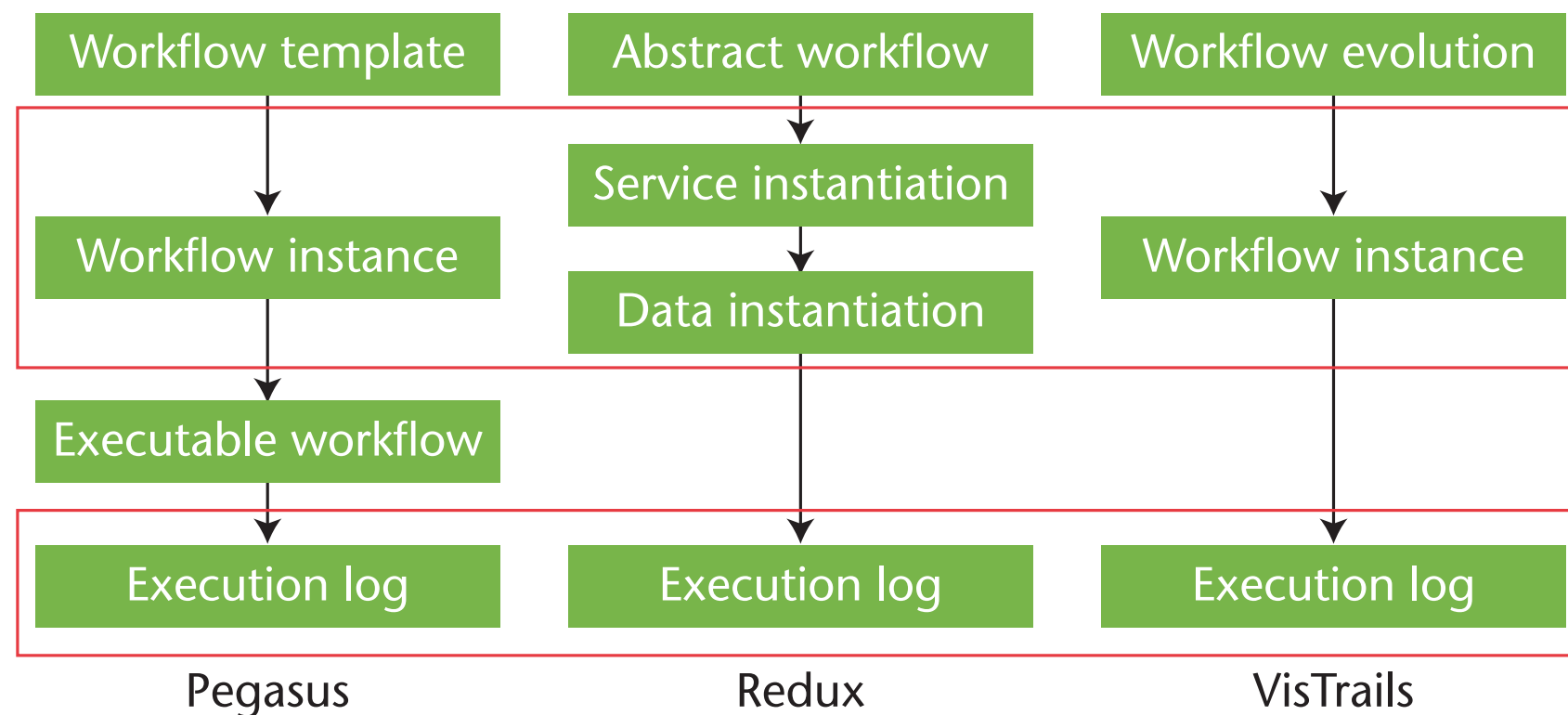
# Provenance Storage Formats

---

- Files, relational databases, XML databases, RDF (linked data)
- Log files are good for preserving data but can be bad to query or analyze
- Relational databases are great for column-specific queries but can be bad for dependency queries
- XML databases are more portable than relational databases but are usually less efficient for queries
- RDF triples are better for dependencies and integrating domain-specific knowledge but can be slower

# Layered Provenance

- As with relational databases, want to normalize provenance to **minimize redundant information**
- Example: Don't store workflow specification each time that workflow is executed—store it once and reference it
- Also allow different layers for different aspects of provenance



[Freire et. al, 2008]



# Provenance Models

---

- How provenance is represented (more abstract than the details of how it is actually stored)
- PROV (W3C Standard) has different storage backends for provenance but all of it conforms to the same model
- Model the objects involved and their relationships (e.g. activities, dependencies)
- Interoperability is a concern
  - Why? May use multiple tools/techniques to achieve a result, want to analyze the entire provenance chain

# Prospective and Retrospective Provenance

---

- Prospective provenance is what was specified/intended
  - a workflow, script, list of steps
- Retrospective provenance is what actually happened
  - actual data, actual parameters, errors that occurred, timestamps, machine information
- **Do not need** prospective provenance to have retrospective provenance!
- Retrospective provenance is often the same type of information as prospective plus more
- Could have multiple retrospective provenance traces for one prospective provenance listing

# Prospective and Retrospective Provenance

---

- **Example:** Baking a Cake
- Prospective Provenance (Recipe):
  1. Gather ingredients (3/4 cup butter, 3/4 cocoa, 3/4 cup flour, ...)
  2. Preheat oven to 350 degrees
  3. Grease cake pan
  4. Mix wet ingredients in large bowl
  5. Mix dry ingredients in a separate bowl
  6. Add dry mixture to wet mixture
  7. Pour batter into cake pan
  8. Put pan in the oven and bake for 30 minutes
  9. Take cake out of oven and let it cool



# Prospective and Retrospective Provenance

---

- Retrospective Provenance (What actually happened)

## 1. Went to store to buy butter

↕ 2. Gathered ingredients (3/4 cup butter, 3/4 cocoa, **1 cup flour**, ...)

3. Greased cake pan

4. Preheated oven to 350 degrees

5. Mixed wet ingredients in large bowl

6. Mixed dry ingredients in a separate bowl

7. Added **wet** mixture to **dry** mixture

8. Poured batter into cake pan

9. Put pan in the oven and baked for **35 minutes**

10. Took cake out of oven and let it cool for **10 minutes**





# Provenance Model History

---

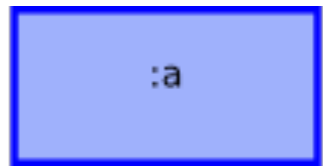
- Community organized provenance challenges (2006-2009)
- First Provenance Challenge assessed capabilities of systems
- Second Provenance Challenge examined interoperability
- Led to development of Open Provenance Model (OPM), (2007)
  - Sought to establish interchange format for provenance
- Further work led to PROV W3C Recommendations (2013)
  - Some confusion from name changes from OPM to PROV even though concepts are similar
  - Focus is on **model** not formats

# PROV: Three Key Classes

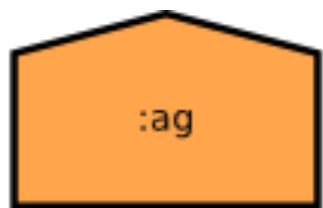
---



An **entity** is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary.



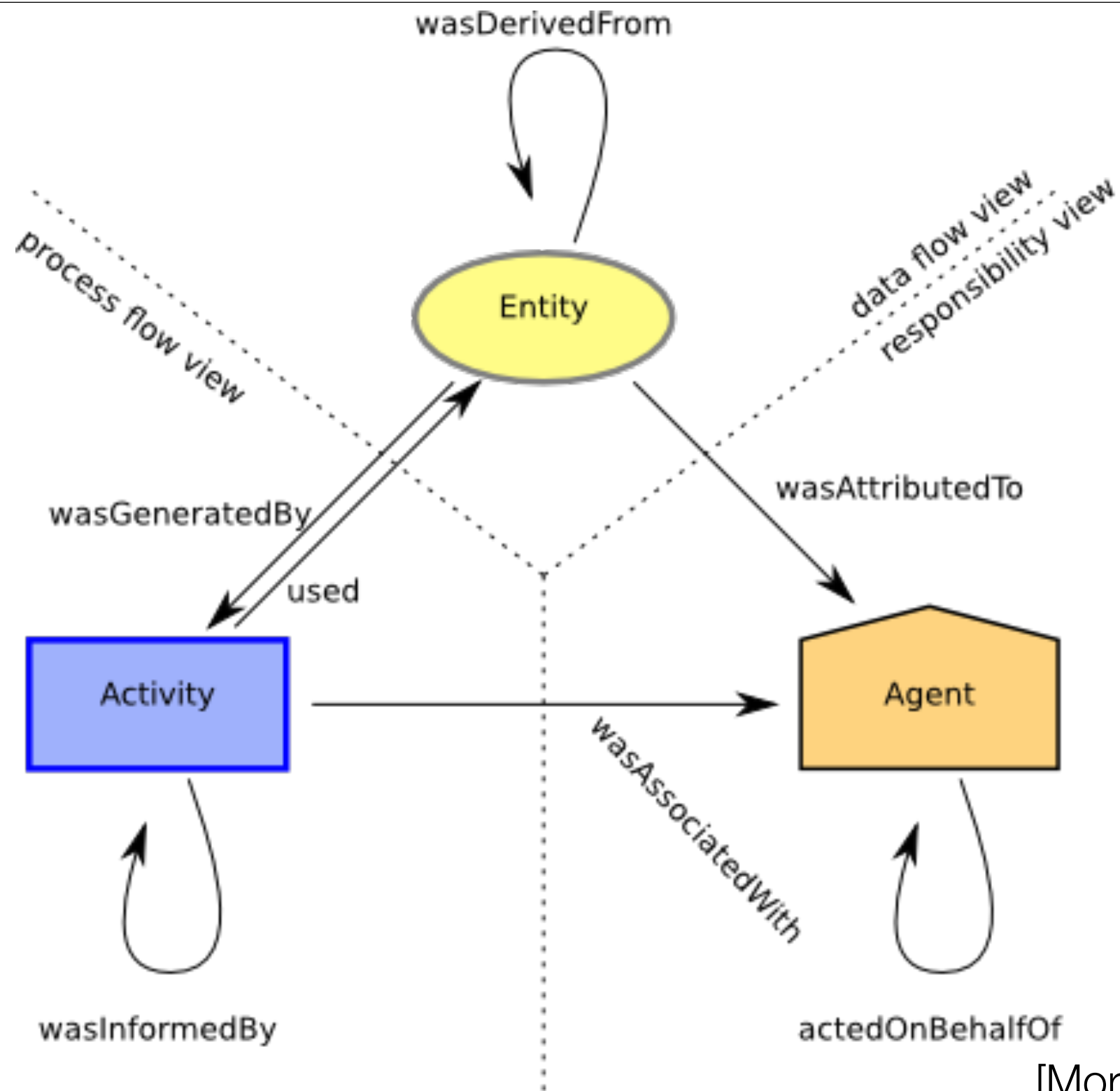
An **activity** is something that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities.



An **agent** is something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity.

[Moreau et al., 2014]

# PROV: Three Views of Provenance



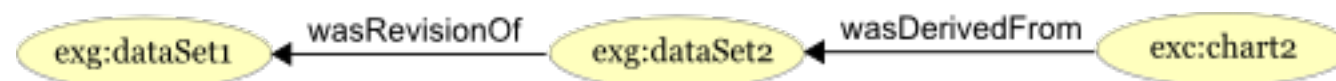
[Moreau et al., 2014]

# PROV Edges: Derivation

- Derivation Edges:
  - wasGeneratedBy: entity  $\rightarrow$  activity
  - used: activity  $\rightarrow$  entity



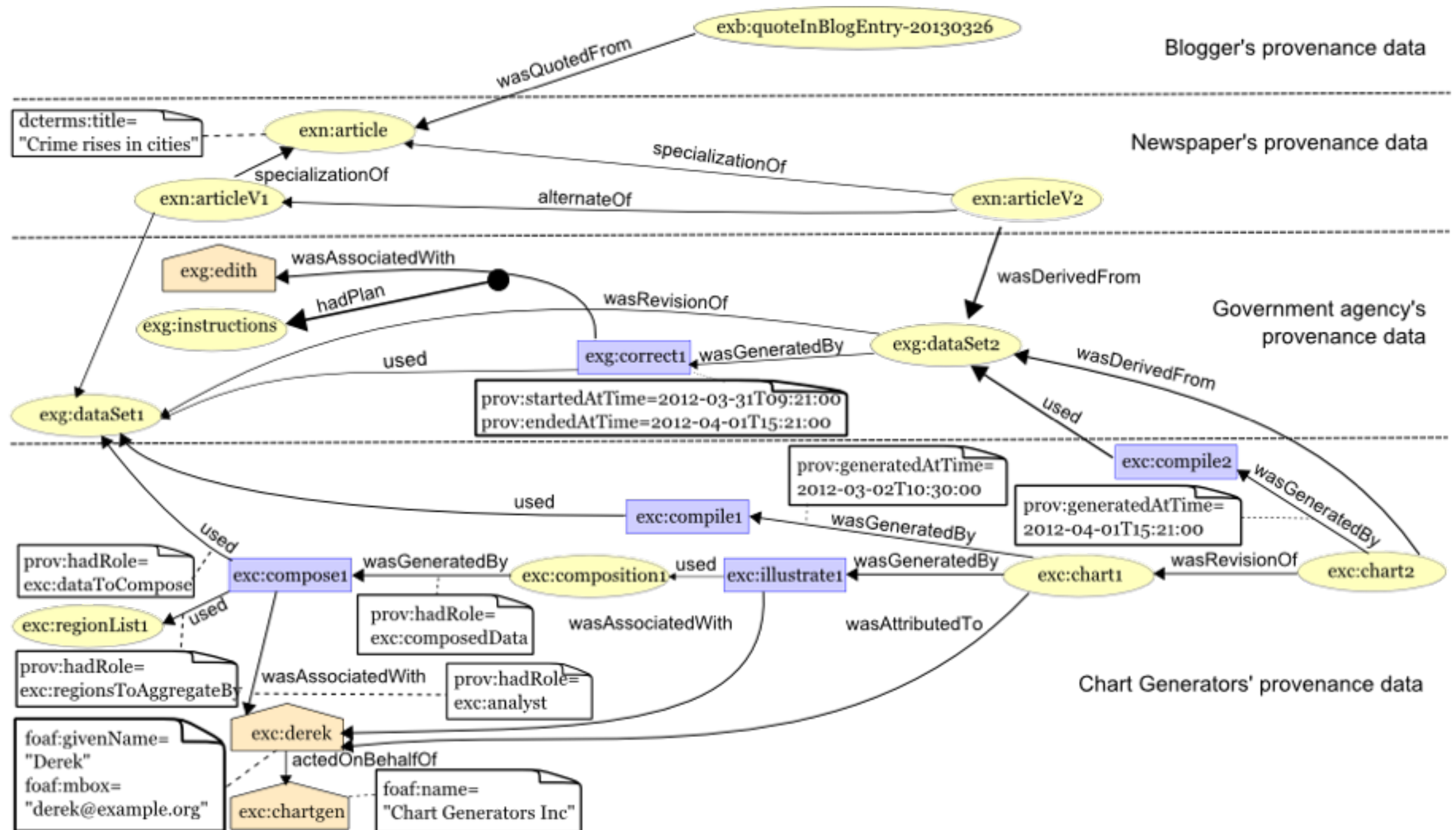
- wasDerivedFrom: entity  $\rightarrow$  entity



[PROV Model Primer, 2013]



# PROV Example



[PROV Model Primer, 2013]

# Querying Provenance

- Query methods are often tied to storage backend
- SQL, XQuery, Prolog, SPARQL, ...

## REDUX

```
SELECT Execution.ExecutableWorkflowId, Execution.ExecutionId, Event.EventId, ExecutableActivity.ExecutableActivityId
from Execution, Execution_Event, Event, ExecutableWorkflow_ExecutableActivity, ExecutableActivity,
     ExecutableActivity_Property_Value, Value, EventType as ET
where Execution.ExecutionId=Execution_Event.ExecutionId
and Execution_Event.EventId=Event.EventId
and ExecutableActivity.ExecutableActivityId=ExecutableActivity_Property_Value.ExecutableActivityId
and ExecutableActivity_Property_Value.ValueId=Value.ValueId and Value.Value=Cast('-m 12' as binary)
and ((CONVERT(DECIMAL, Event.Timestamp)+0)%7)=0 and Execution_Event.ExecutableWorkflow_ExecutableActivityId=
     ExecutableWorkflow_ExecutableActivity.ExecutableWorkflow_ExecutableActivityId
and ExecutableWorkflow_ExecutableActivity.ExecutableWorkflowId=Execution.ExecutableWorkflowId
and ExecutableWorkflow_ExecutableActivity.ExecutableActivityId=ExecutableActivity.ExecutableActivityId
and Event.EventType=ET.EventType and ET.EventTypeName='Activity Start';
```

## VisTrails

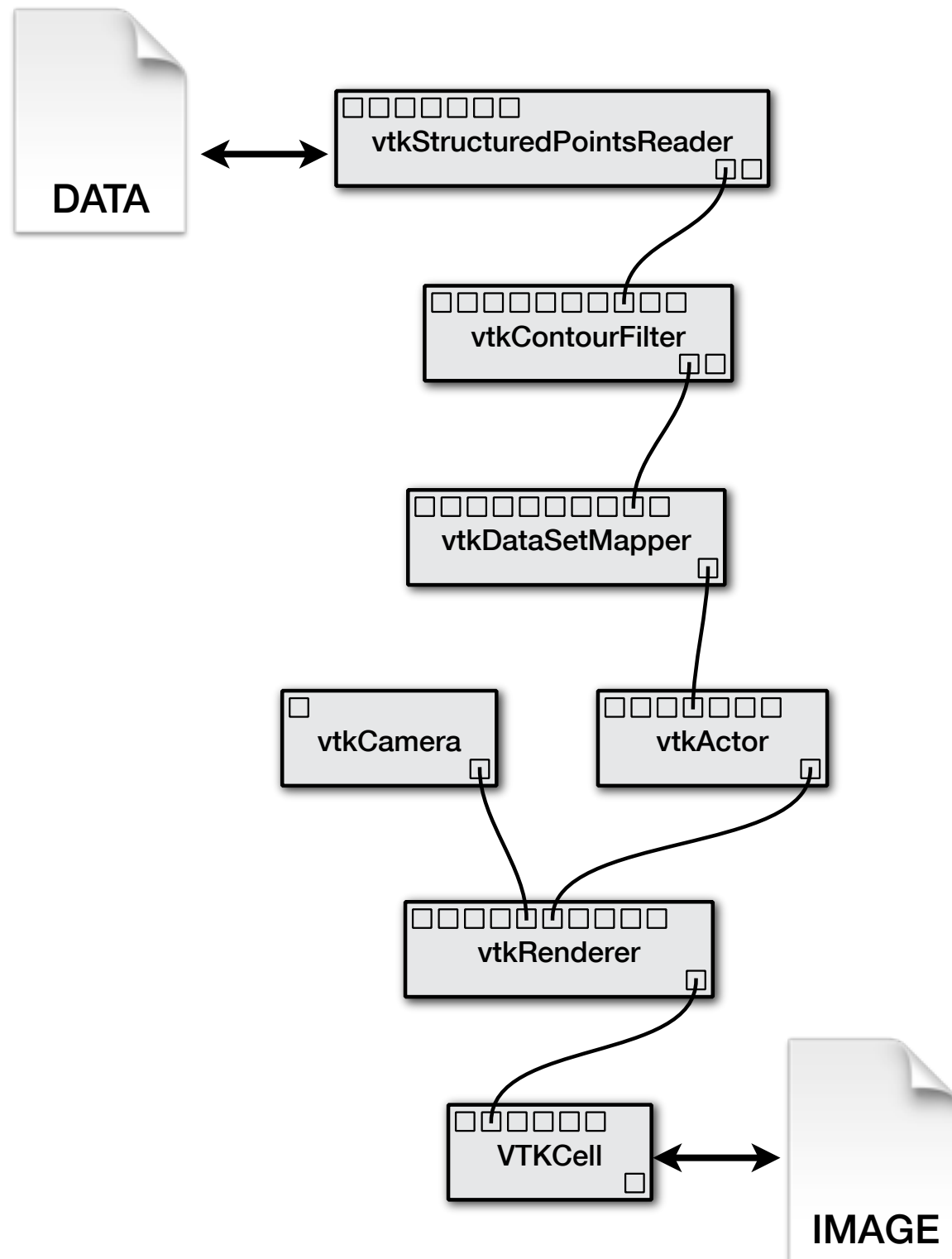
```
wf{*}: x where x.module='AlignWarp' and x.parameter('model')='12'
and (log{x}: y where y.dayOfWeek='Monday')
```

## MyGrid

```
SELECT ?p
where (?p <http://www.mygrid.org.uk/provenance#startTime> ?time) and (?time > date)
using ns for <http://www.mygrid.org.uk/provenance#> xsd for <http://www.w3.org/2001/XMLSchema#>

SELECT ?p
where <urn:lsid:www.mygrid.org.uk:experimentinstance:HXQOVQA2ZI0>
(?p <http://www.mygrid.org.uk/provenance#runsProcess> ?processname .
?p <http://www.mygrid.org.uk/provenance#processInput> ?inputParameter .
?inputParameter <ont:model> <ontology:twelfthOrder>)
using ns for <http://www.mygrid.org.uk/provenance#> ont for <http://www.mygrid.org.uk/ontology#>
```

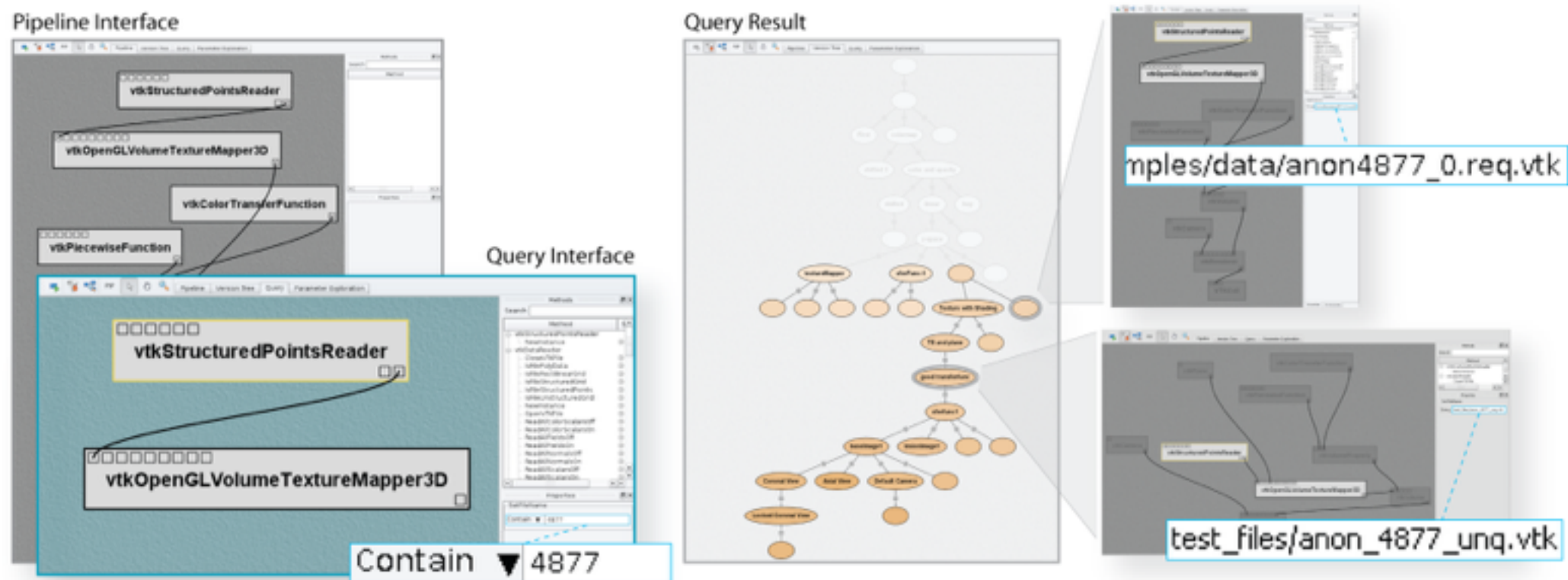
# Querying Provenance



- *What process led to the output image?*
- *What input datasets contributed to the output image?*
- *What workflows include resampling and isosurfacing with isovalue 57?*
- Graph traversal or graph patterns
  - How do we write such queries?

# Querying Provenance by Example

- Provenance is represented as graphs: hard to specify queries using text!
- Querying workflows by example [Scheidegger et al., TVCG 2007; Beerli et al., VLDB 2006; Beerli et al. VLDB 2007]
  - WYSIWYQ -- What You See Is What You Query
  - Interface to create workflow is same as to query





# Stronger Links Between Provenance and Data

```
<workflow_exec id="1">
  <m_exec id="5"
    name="vtkStructuredDataReader"
    package="edu.utah.sci.vistrails.vtk"
    version="5.6.0">
    <param id="2" name="SetFile"
      value="/MyData/05-12-sc2.dat"/>
  </m_exec>
  <m_exec id="6"
    name="vtkContourFilter"
    package="edu.utah.sci.vistrails.vtk"
    version="5.6.0">
    <param id="3" name="SetValue"
      value="[1, 57]"/>
    <param id="4" name="ComputeScalarsOn"
      value="True"/>
  </m_exec>
  ...
  <m_exec id="11"
    name="FileSink"
    package="edu.utah.sci.vistrails.basic"
    version="1.5">
    <param id="15" name="path"
      value="/home/a/results/23.out"/>
  </m_exec>
```

- Filenames are often the mode of identification in data exploration
- We might also use URIs or access curated data stores
  - Can this always be expected for exploratory tasks?
  - What happens if offline?
- Solution:
  - Managed store for data associated with computations
  - Improved data identification
  - Automatic versioning

[Koop et. al, 2010]

# Stronger Links Between Provenance and Data

```
<workflow_exec id="1">
  <m_exec id="5"
    name="vtkStructuredDataReader"
    package="edu.utah.sci.vistrails.vtk"
    version="5.6.0">
    <param id="2" name="SetFile"
      value="/MyData/05-12-s
  </m_exec>
  <m_exec id="6"
    name="vtkContourFilter"
    package="edu.utah.sci.vistrails.vtk"
    version="5.6.0">
    <param id="3" name="SetValue"
      value="[1, 57]"/>
    <param id="4" name="ComputeScalarsOn"
      value="True"/>
  </m_exec>
  ...
  <m_exec id="11"
    name="FileSink"
    package="edu.utah.sci.vistrails.basic"
    version="1.5">
    <param id="15" name="path"
      value="/home/a/results/23.out"/>
  </m_exec>
```



**FILE NOT FOUND**

- Filenames are often the mode of identification in data exploration
- We might also use URIs or access curated data stores
  - Can this always be expected for exploratory tasks?
  - What happens if offline?
- Solution:
  - Managed store for data associated with computations
  - Improved data identification
  - Automatic versioning

[Koop et. al, 2010]

# Stronger Links Between Provenance and Data

```
<workflow_exec id="1">
  <m_exec id="5"
    name="vtkStructuredDataReader"
    package="edu.utah.sci.vistrails.vtk"
    version="5.6.0">
    <param id="2" name="SetFile"
      value="/MyData/05-12-s
    </m_exec>
    <m_exec id="6"
      name="vtkContourFilter"
      package="edu.utah.sci.vistrails.vtk"
      version="5.6.0">
      <param id="3" name="SetValue"
        value="[1, 57]"/>
      <param id="4" name="ComputeScalarsOn"
        value="True"/>
      </m_exec>
      ...
      <m_exec id="11"
        name="FileSink"
        package="edu.utah.sci.vistrails.basic"
        version="1.5">
        <param id="15" name="path"
          value="/home/a/results
        </m_exec>
```



**FILE NOT FOUND**

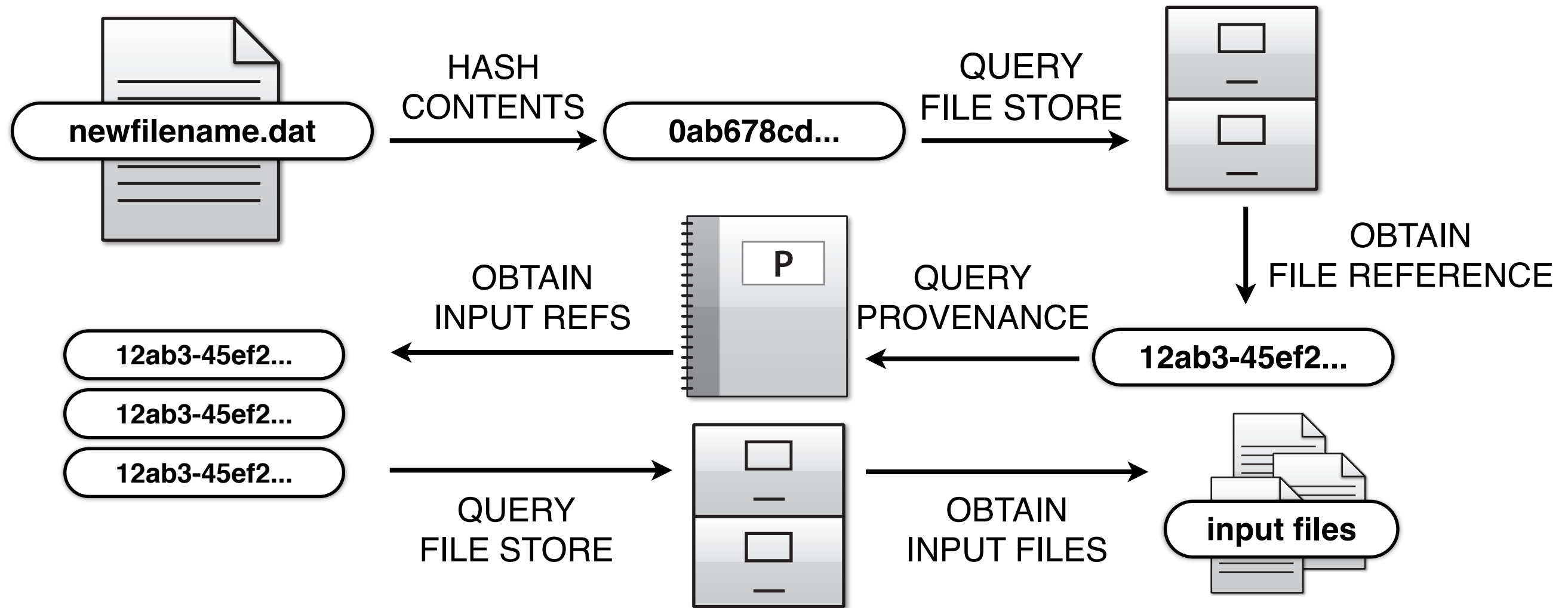


**FILE NOT FOUND**

- Filenames are often the mode of identification in data exploration
- We might also use URIs or access curated data stores
  - Can this always be expected for exploratory tasks?
  - What happens if offline?
- Solution:
  - Managed store for data associated with computations
  - Improved data identification
  - Automatic versioning

[Koop et. al, 2010]

# Provenance from Data



[Koop et. al, 2010]

# Provenance-Enabled Systems

**Table 1. Provenance-enabled systems.**

System	Capture mechanism	Prospective provenance	Retrospective provenance	Workflow evolution
REDUX	Workflow-based	Relational	Relational	No
Swift	Workflow-based	SwiftScript	Relational	No
VisTrails	Workflow-based	XML and relational	Relational	Yes
Karma	Workflow- and process-based	Business Process Execution Language	XML	No
Kepler	Workflow-based	MoML	MoML variation	Under development
Taverna	Workflow-based	Scufl	RDF	Under development
Pegasus	Workflow-based	OWL	Relational	No
PASS	OS-based	N/A	Relational	No
ES3	OS-based	N/A	XML	No
PASOA/PreServ	Process-based	N/A	XML	No

[Freire et. al, 2008]



# Provenance-Enabled Systems

More...

**Table 1. Provenanc**

System	Storage	Query support	Available as open source?
REDUX	Relational database management system (RDBMS)	SQL	No
Swift	RDBMS	SQL	Yes
VisTrails	RDBMS and files	Visual query by example, specialized language	Yes
Karma	RDBMS	Proprietary API	Yes
Kepler	Files; RDBMS planned	Under development	Yes
Taverna	RDBMS	SPARQL	Yes
Pegasus	RDBMS	SPARQL for metadata and workflow; SQL for execution log	Yes
PASS	Berkeley DB	nq (proprietary query tool)	No
ES3	XML database	XQuery	No
PASOA/PreServ	Filesystem, Berkeley DB	XQuery, Java query API	Yes

[Freire et. al, 2008]

# Provenance-Enabled Systems



*Sumatra*

IPython Notebook

More...

Table 1. Provenanc

System	Storage	Query support	Available as open source?
REDUX	Relational database management system (RDBMS)	SQL	No
Swift	RDBMS	SQL	Yes
VisTrails	RDBMS and files	Visual query by example, specialized language	Yes
Karma	RDBMS	Proprietary API	Yes
Kepler	Files; RDBMS planned	Under development	Yes
Taverna	RDBMS	SPARQL	Yes
Pegasus	RDBMS	SPARQL for metadata and workflow; SQL for execution log	Yes
PASS	Berkeley DB	nq (proprietary query tool)	No
ES3	XML database	XQuery	No
PASOA/PreServ	Filesystem, Berkeley DB	XQuery, Java query API	Yes

[Freire et. al, 2008]

# VisTrails

---



- Comprehensive provenance infrastructure for computational tasks
- Focus on exploratory tasks such as simulation, visualization, and data analysis
- Transparently tracks provenance of the discovery process—from data acquisition to visualization
  - The trail followed as users generate and test hypotheses
  - Users can refer back to any point along this trail at any time
- Leverage provenance to streamline exploration
- Focus on usability—build tools for scientists

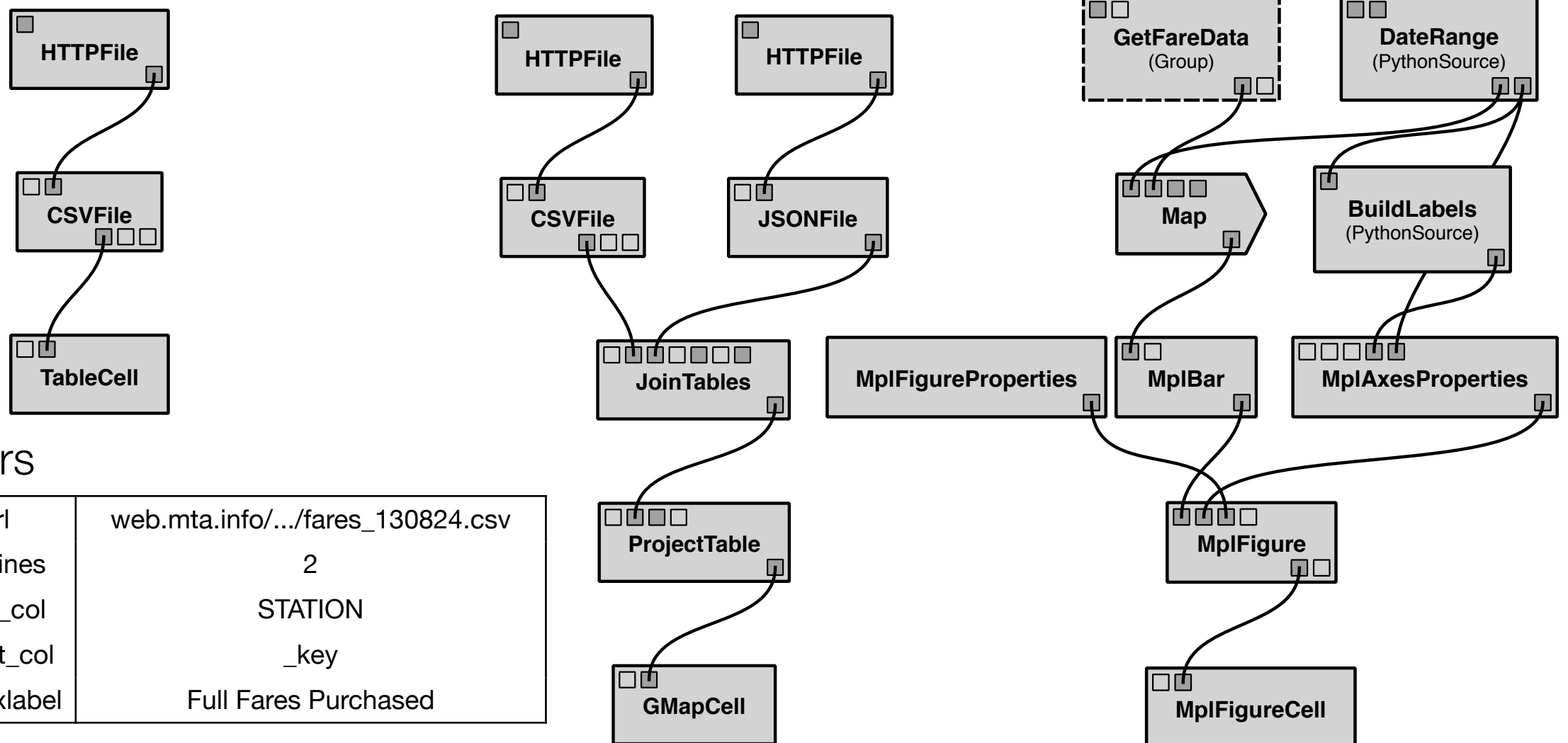
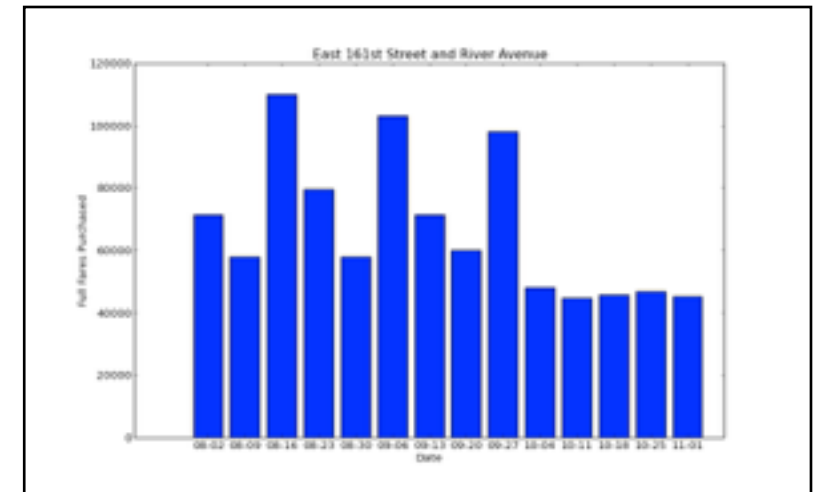
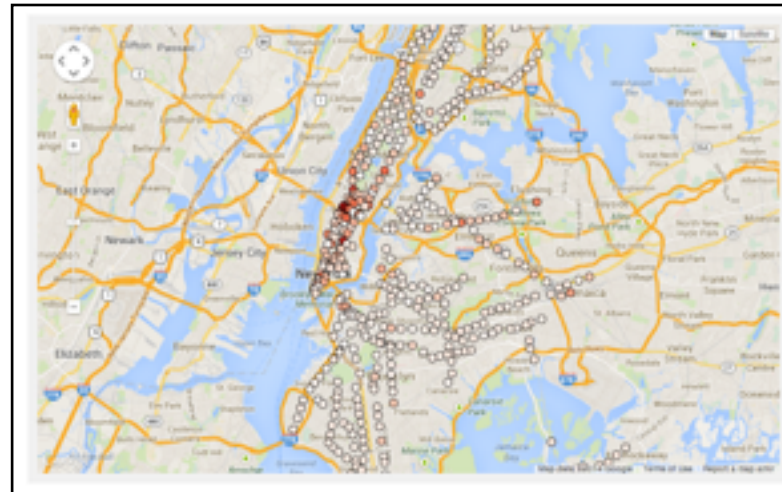
# VisTrails



- Open-source, freely downloadable system ([www.vistrails.org](http://www.vistrails.org))
  - Also on GitHub ([github.com/vistrails](https://github.com/vistrails)), PyPI, conda package
- Multi-platform: users on Mac, Linux, and Windows
- Python code and uses PyQt and Qt for the interface
- Over 35,000 downloads
- User's guide, wiki, and mailing list
- Many users in different disciplines and countries:
  - Visualizing environmental simulations (CMOP STC)
  - Simulation for solid, fluid and structural mechanics (Galileo Network, UFRJ Brazil)
  - Quantum physics simulations (ALPS, ETH Zurich)
  - Climate analysis (UV-CDAT, LLNL)
  - Habitat modeling (USGS)
  - Open Wildland Fire Modeling (U. Colorado, NCAR)
  - High-energy physics (LEPP, Cornell)
  - Cosmology simulations (LANL)
  - Using tms for improving memory (Psychiatry, U. Utah)
  - eBird (Cornell, NSF DataONE)
  - Astrophysical Systems (LSU)
  - NIH NBCR (UCSD)
  - Pervasive Technology Labs (Indiana University)
  - Linköping University
  - University of North Carolina, Chapel Hill
  - UTEP

# Example: Workflows

ROUTE	STATION	RT	Y	SECT	7-D-AD-UN	2-AD-UN	1-D-UN	7-D-UN	2-D-UN
1	42ND STREET & 8TH AVENUE	00228885	00008475	00000441	00000455	00000034	00003341	00071255	
2	14TH STREET-UNION SQUARE	00224603	00013051	00000627	00000026	00000660	00089167	00199841	
3	42ND STREET & GRAND CENTRAL	00207718	00007908	00000123	00000183	00003801	00048759	00096603	
4	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00000279	00003822	00010127	00067483	
5	34TH STREET - PENN STATION	00168768	00006155	00000123	00000065	00000831	00030645	00054376	
6	42ND STREET/TIMES SQUARE	00159382	00005945	00000178	00000205	00000699	00018931	00078644	
7	34TH STREET & 6TH AVENUE	00156808	00006276	00000487	00000143	00000712	00018920	00120486	
8	19TH STREET/COLUMBUS CIRCLE	00151262	00009484	00000189	00000071	00000542	00013187	00119966	
9	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000184	00000159	00000723	00037978	00090745	
10	86TH STREET-LEXINGTON AVE	00142369	00010067	00000470	00000839	00000275	00010328	00125250	
11	34TH STREET & 6TH AVENUE	00134812	00009005	00000348	00000112	00000649	00011131	00075040	
12	PARK PLACE	00121814	00004311	00000287	00000931	00000792	00025484	00065362	
13	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000784	00000241	00022808	00068256	
14	34TH STREET & 7TH AVENUE	00091876	00003990	00000232	00000727	00000459	00024284	00038671	
15	LEXINGTON AVENUE	00084815	00004688	00000190	00000833	00000754	00020018	00015066	
16	8TH AVENUE-34TH STREET	00084313	00003907	00000286	00000144	00000236	00018272	00076661	
17	BARCLAYS CENTER	00083804	00004204	00000454	00000186	00000495	00018113	00068119	
18	WFLY 4TH ST-WASHINGTON ST	00081562	00004677	00000751	00000965	00000177	00011628	00074458	



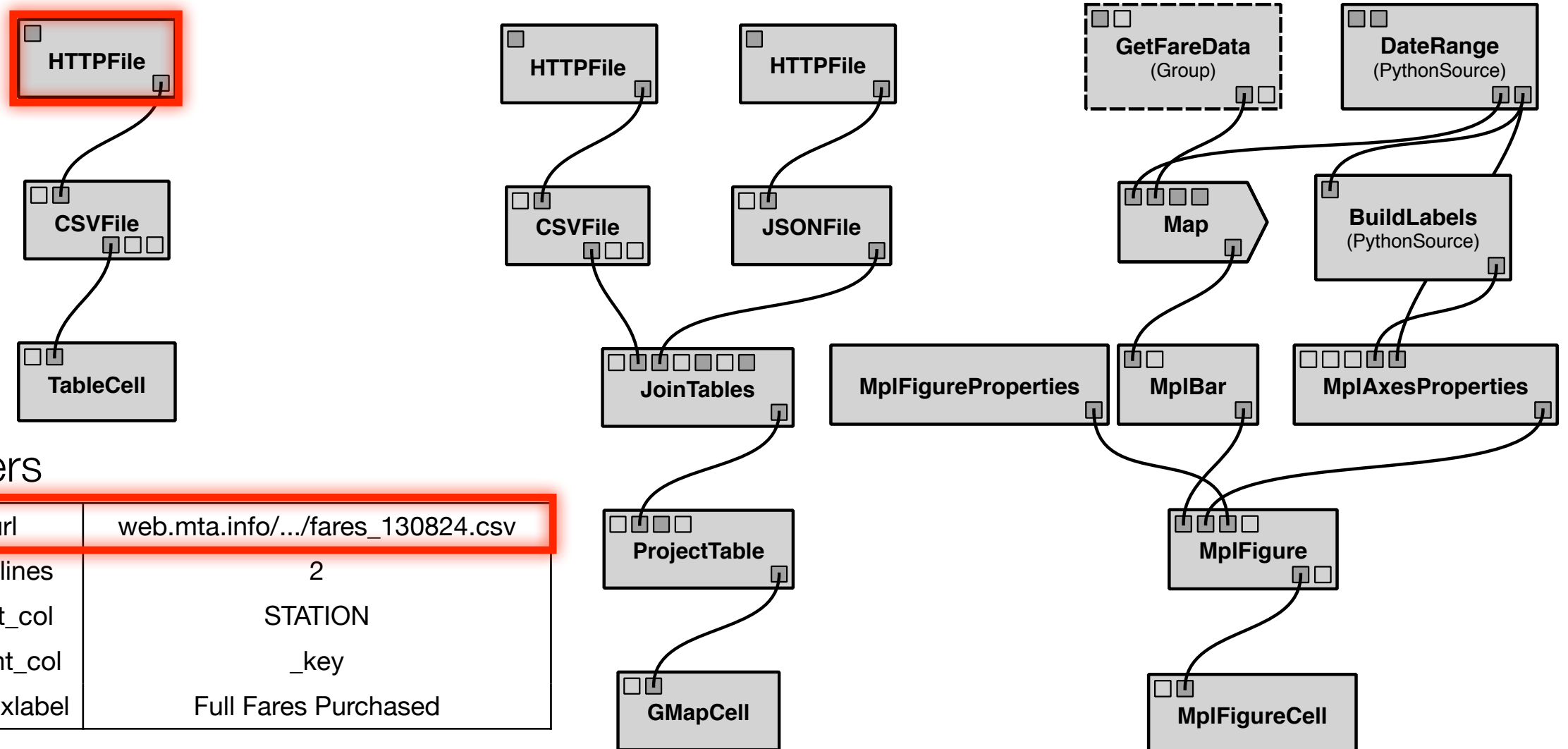
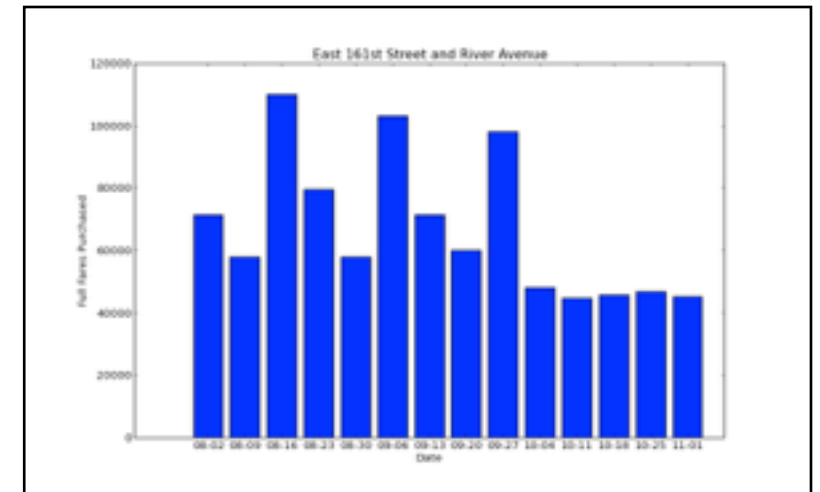
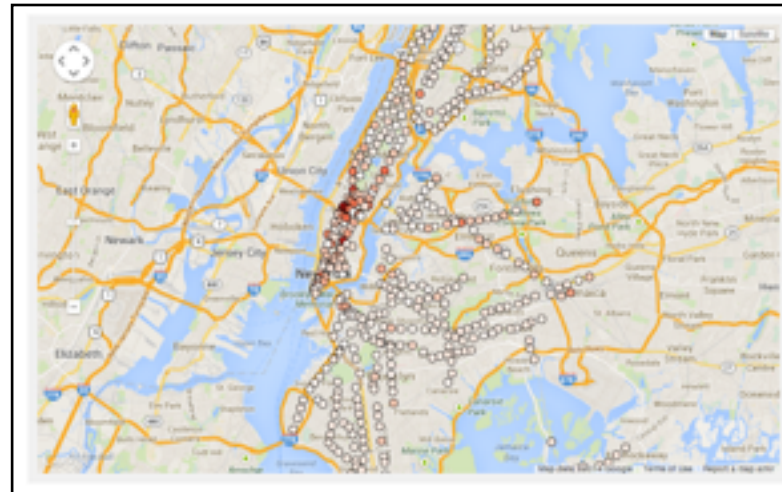
## Parameters

HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased



# Example: Workflows

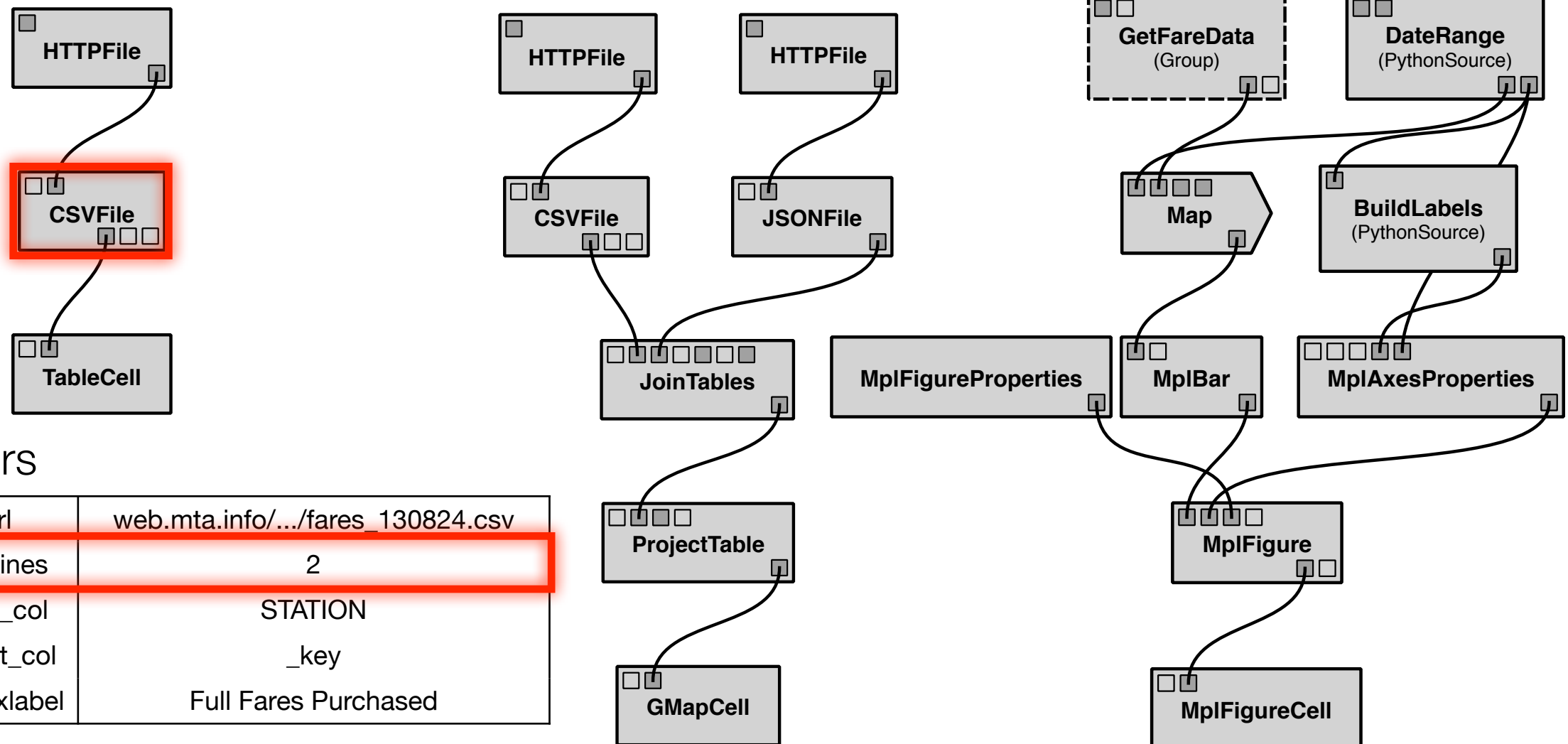
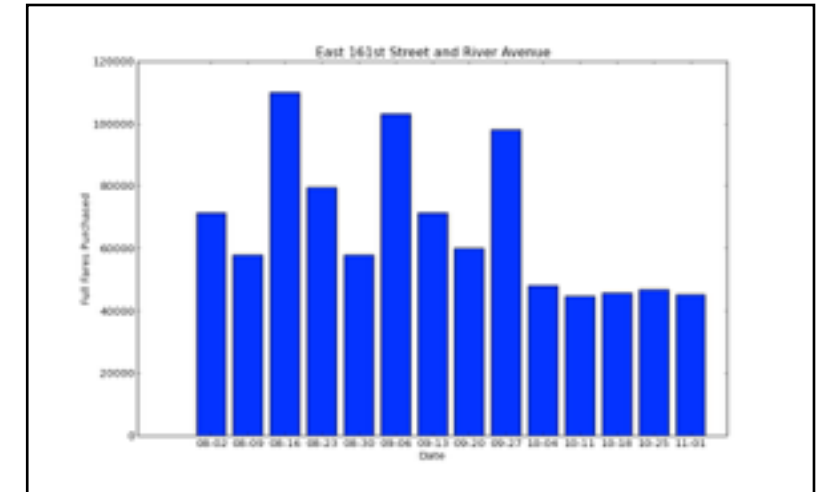
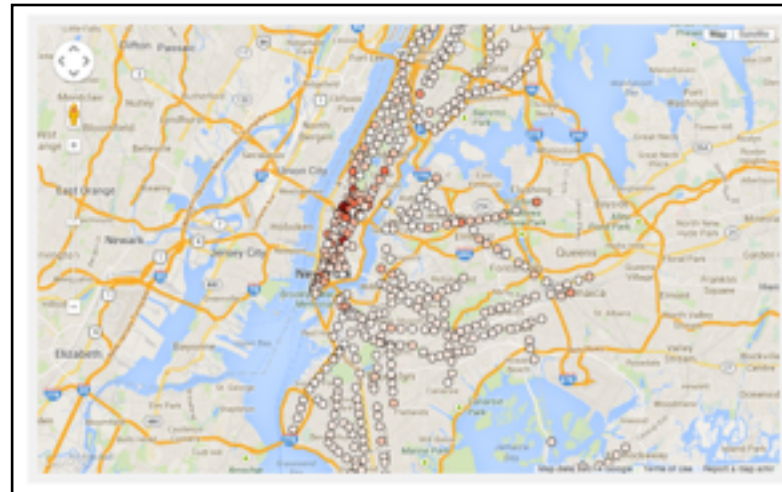
ROUTE	STATION	RT	Y	SECT	7-D-AD-UN	2-AD-UN	1-D-UN	7-D-UN	2-D-UN
1	42ND STREET & 8TH AVENUE	00228885	00008475	00000441	00000455	00000034	00003341	00071255	
2	14TH STREET-UNION SQUARE	00224603	00013051	00000627	00003026	00000660	00089367	00199841	
3	42ND STREET & GRAND CENTRAL	00207718	00007908	00000123	00001183	00003801	00048759	00096603	
4	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00001279	00003822	00010127	00067483	
5	34TH STREET - PENN STATION	00168768	00006155	00000123	00000965	00000831	00030645	00054376	
6	42ND STREET/TIMES SQUARE	00159382	00005945	00000178	00001205	00000699	00018931	00078644	
7	34TH STREET & 6TH AVENUE	00156808	00006276	00000487	00001143	00000712	00018920	00120486	
8	19TH STREET/COLUMBUS CIRCLE	00151262	00000484	00000189	00002071	00000542	00013187	00119966	
9	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000184	00001159	00000723	00037978	00090745	
10	86TH STREET-LEXINGTON AVE	00142589	00010967	00000470	00000839	00000275	00018328	00125250	
11	34TH STREET & 6TH AVENUE	00134812	00009005	00000348	00001112	00000649	00011131	00075040	
12	PARK PLACE	00121814	00004311	00000287	00000931	00000792	00025484	00065362	
13	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000784	00001241	00022808	00068256	
14	34TH STREET & 7TH AVENUE	00091876	00003990	00000232	00000727	00001459	00024284	00038671	
15	LEXINGTON AVENUE	00084815	00004688	00000190	00000833	00000754	00020018	00015066	
16	8TH AVENUE-34TH STREET	00084313	00003907	00000286	00001144	00000236	00018272	00076661	
17	BARCLAYS CENTER	00083804	00004204	00000454	00001186	00001495	00018113	00068119	
18	WFLY 4TH ST-WASHINGTON ST	00081562	00004677	00000751	00000965	00000177	00011628	00074458	



## Parameters

HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

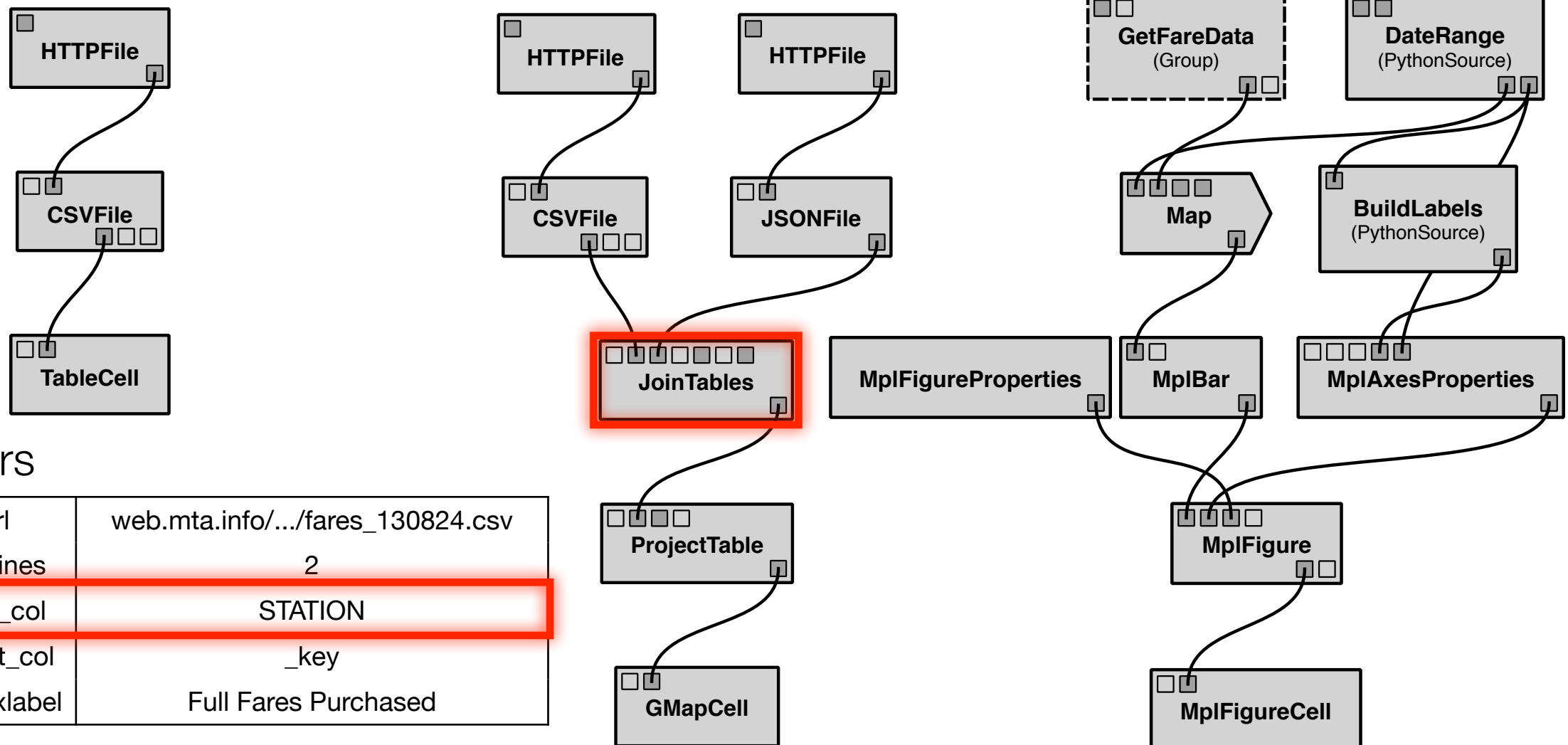
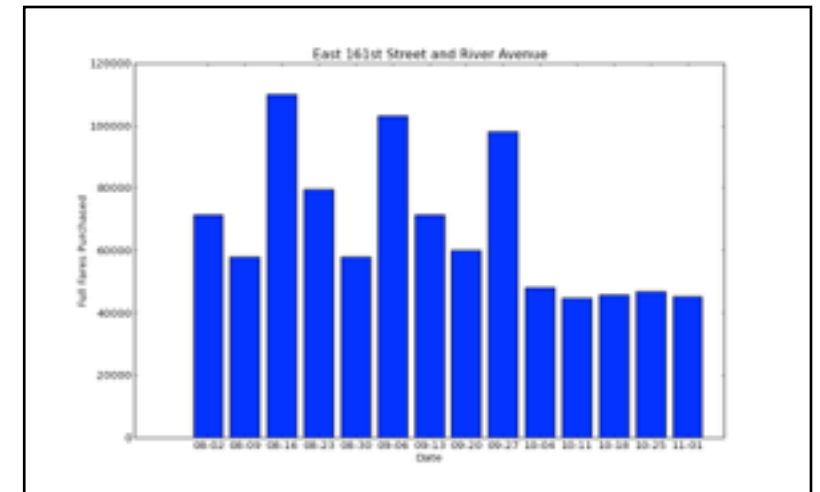
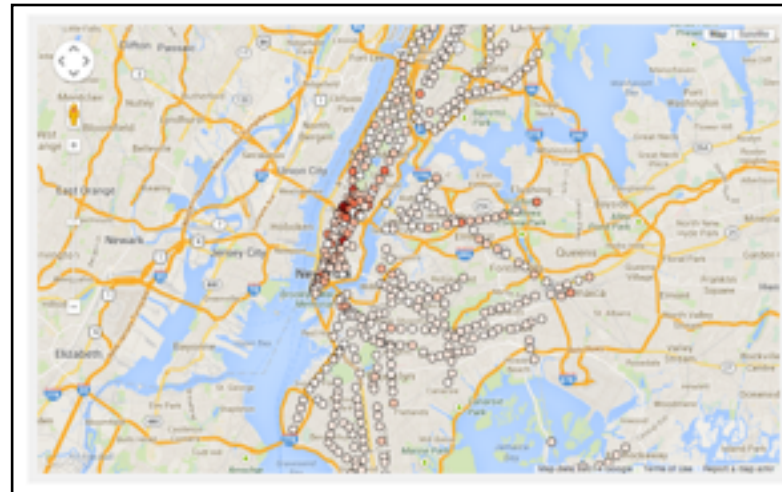
	ROUTE	STATION	FF	Y	SEN/DE	P-D-AM-UN	P-D-AM-UNP	JOINT BA TXT	P-D-UNL	JO-D-UNL
1	R011	42ND STREET & 8TH AVENUE	00228895	00008471	00000441	00000451	00000134	00013141	00071255	
2	R170	34TH STREET-UNION SQUARE	00324603	00011051	00000627	00003026	00000660	00089167	00109841	
3	R046	42ND STREET & GRAND CENTRAL	00307718	00007968	00000123	00000183	00003865	00048759	00094613	
4	R012	34TH STREET & 8TH AVENUE	00188311	00006690	00000498	00001279	00001822	00035127	00067483	
5	R290	34TH STREET - PENN STATION	00168768	00006155	00000123	00000065	00000831	00010645	00014376	
6	R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000178	00000205	00000690	00018931	00078644	
7	R022	34TH STREET & 6TH AVENUE	00166008	00006276	00000487	00000143	00000712	00018920	00120466	
8	R084	59TH STREET/COLUMBUS CIRCLE	00151262	00009484	00000189	00002071	00000542	00019187	00119966	
9	R020	47-50 STREETS/ROCKEFELLER	00143509	00006402	00000184	00000159	00000673	00037978	00090745	
10	R179	86TH STREET-LEXINGTON AVE	00145268	00010367	00000470	00000839	00000271	00016128	00125250	
11	R023	34TH STREET & 6TH AVENUE	00134012	00005905	00000148	00000112	00000649	00011131	00075040	
12	R029	PARK PLACE	00121614	00004311	00000187	00000931	00000792	00025406	00061362	
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000704	00001243	00022888	00068216	
14	R031	34TH STREET & 7TH AVENUE	00095876	00003990	00000232	00000727	00001419	00024284	00018671	
15	R017	LEXINGTON AVENUE	00094615	00004688	00000190	00000833	00000714	00020018	00015066	
16	R175	8TH AVENUE-14TH STREET	00094313	00003907	00000286	00000144	00000216	00018272	00074681	
17	R017	BARCLAYS CENTER	00093804	00004204	00000454	00000186	00001490	00039113	00068119	
18	R118	WFL 4TH ST - WASHINGTON SQ	00091562	00004677	00000351	00000961	00000177	00011628	00074418	



HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

# Example: Workflows

ROUTE	STATION	RT	T	SEV/DIS	F-D-AM-UN	D-AM-UN	J-ORF-RR-TAT	F-D-UN	D-D-UN
1 R011	42ND STREET & 8TH AVENUE	00228885	00008475	00000441	00000455	00000034	00003341	00071255	
2 R170	34TH STREET-UNION SQUARE	00224603	00013051	00000627	00000026	00000660	00089167	00199841	
3 R046	42ND STREET & GRAND CENTRAL	00207718	00007908	00000123	00000183	00003801	00048759	00096603	
4 R012	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00000279	00003822	00010127	00067483	
5 R295	34TH STREET - PENN STATION	00168768	00006155	00000123	00000065	00000831	00030645	00054376	
6 R033	42ND STREET/TIMES SQUARE	00159382	00005945	00000178	00000205	00000699	00018931	00078644	
7 R022	34TH STREET & 6TH AVENUE	00156808	00006276	00000487	00000143	00000712	00018920	00120486	
8 R086	19TH STREET/COLUMBUS CIRCLE	00151262	00009484	00000189	00000207	00000542	00013187	00119966	
9 R039	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000184	00000159	00000723	00037978	00090745	
10 R179	86TH STREET-LEXINGTON AVE	00142369	00010067	00000470	00000839	00000275	00018328	00125250	
11 R023	34TH STREET & 6TH AVENUE	00134812	00009005	00000348	00000112	00000649	00010131	00075040	
12 R029	PARK PLACE	00121814	00004311	00000287	00000931	00000792	00025484	00065362	
13 R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000784	00000241	00022888	00068256	
14 R031	34TH STREET & 7TH AVENUE	00091876	00003990	00000232	00000727	00000459	00024284	00038671	
15 R017	LEXINGTON AVENUE	00084815	00004688	00000190	00000833	00000754	00020018	00015066	
16 R175	8TH AVENUE-34TH STREET	00084313	00003907	00000286	00000144	00000236	00018272	00074661	
17 R057	BARCLAYS CENTER	00083804	00004204	00000454	00000186	00000495	00018113	00068119	
18 R116	WFLY 4TH ST-WASHINGTON ST	00081562	00004677	00000751	00000965	00000577	00015428	00074458	



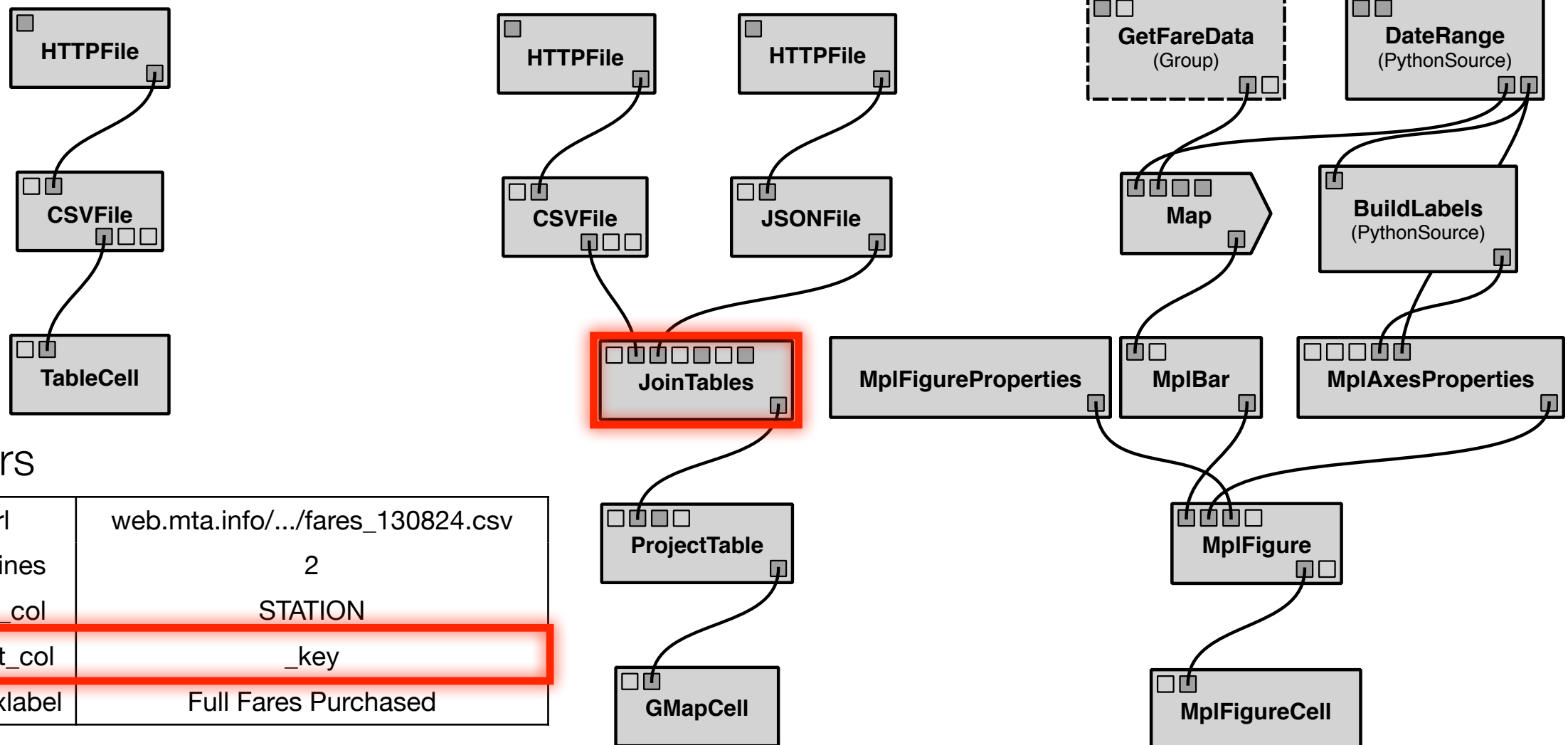
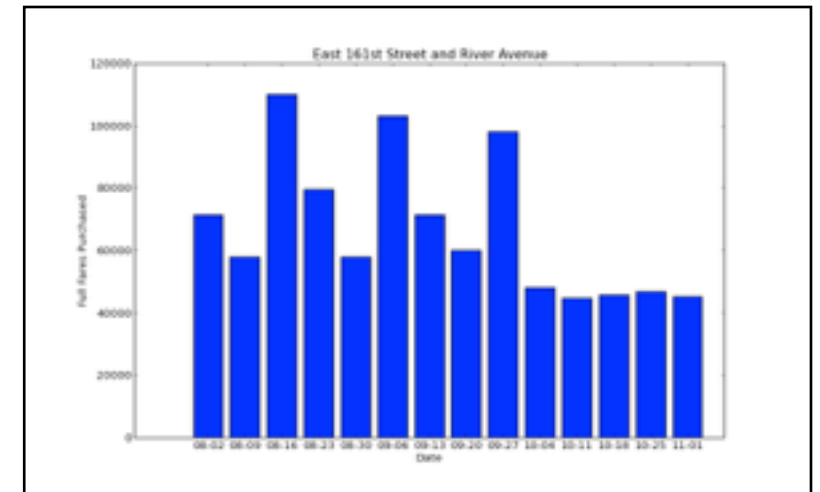
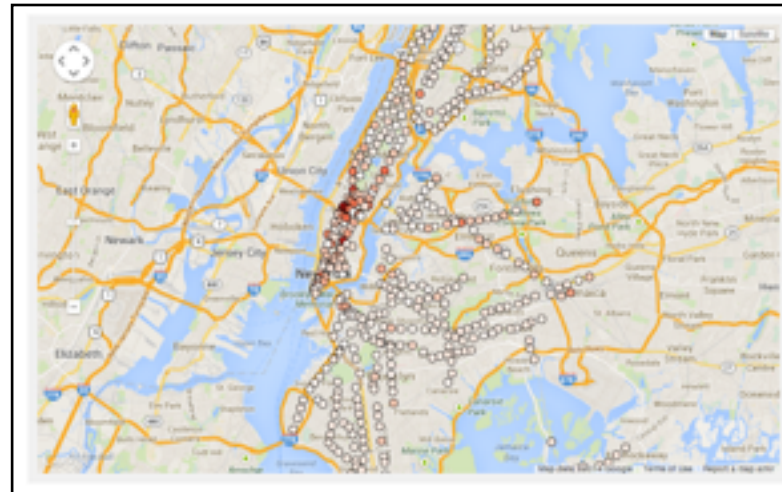
## Parameters

HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased



# Example: Workflows

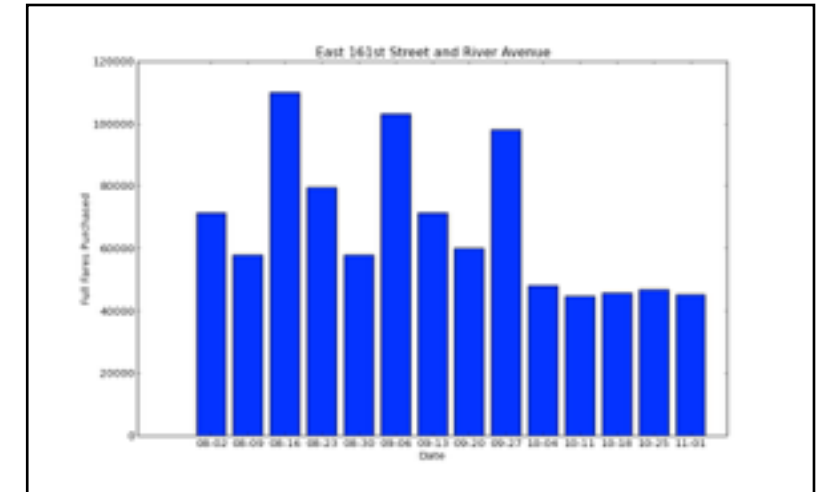
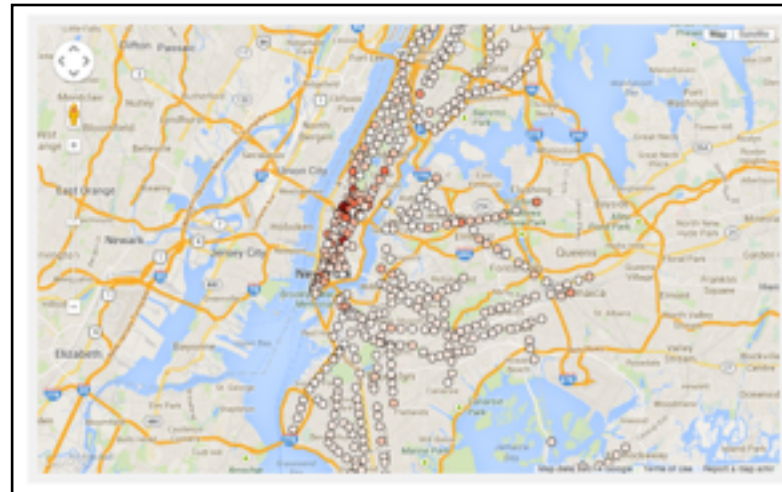
ROUTE	STATION	RT	Y	SENDS	F-D-AS-UN	D-AS-UN	JOINT	RT-T	F-D-UN	D-UN
1	42ND STREET & 8TH AVENUE	00228885	00008475	00000441	00000455	00000034	00003341	00071255		
2	14TH STREET-UNION SQUARE	00224603	00013051	00000627	00000026	00000660	00089167	00199841		
3	42ND STREET & GRAND CENTRAL	00207718	00007908	00000123	00000183	00003801	00048759	00096603		
4	34TH STREET & 8TH AVENUE	00188311	00006490	00000498	00000279	00003822	00010127	00067483		
5	34TH STREET - PENN STATION	00168768	00006155	00000123	00000065	00000831	00030645	00054376		
6	42ND STREET/TIMES SQUARE	00159382	00005945	00000178	00000205	00000699	00018931	00078644		
7	34TH STREET & 6TH AVENUE	00156808	00006276	00000487	00000143	00000712	00018920	00120486		
8	19TH STREET/COLUMBUS CIRCLE	00151262	00009484	00000189	00000207	00000542	00013187	00119966		
9	47-50 STREETS/ROCKEFELLER	00143500	00006402	00000184	00000159	00000723	00037978	00090745		
10	86TH STREET-LEXINGTON AVE	00142389	00010067	00000470	00000839	00000275	00018328	00125250		
11	34TH STREET & 6TH AVENUE	00134812	00009005	00000348	00000112	00000649	00011131	00075040		
12	PARK PLACE	00121814	00004311	00000287	00000931	00000792	00025484	00065362		
13	42ND STREET & GRAND CENTRAL	00100742	00004273	00000185	00000784	00000124	00022808	00068256		
14	34TH STREET & 7TH AVENUE	00091876	00003990	00000232	00000727	00000459	00024284	00038671		
15	LEXINGTON AVENUE	00084815	00004688	00000190	00000833	00000754	00020018	00015066		
16	8TH AVENUE-34TH STREET	00084313	00003907	00000286	00000144	00000236	00018272	00076661		
17	BARCLAYS CENTER	00083804	00004204	00000454	00000186	00000495	00018113	00068119		
18	WFLY 4TH ST-WASHINGTON ST	00081562	00004677	00000751	00000965	00000177	00015428	00074458		



## Parameters

HTTPFile.url	web.mta.info/.../fares_130824.csv
CSVFile.skip_lines	2
JoinTables.left_col	STATION
JoinTables.right_col	_key
MplAxesProps.xlabel	Full Fares Purchased

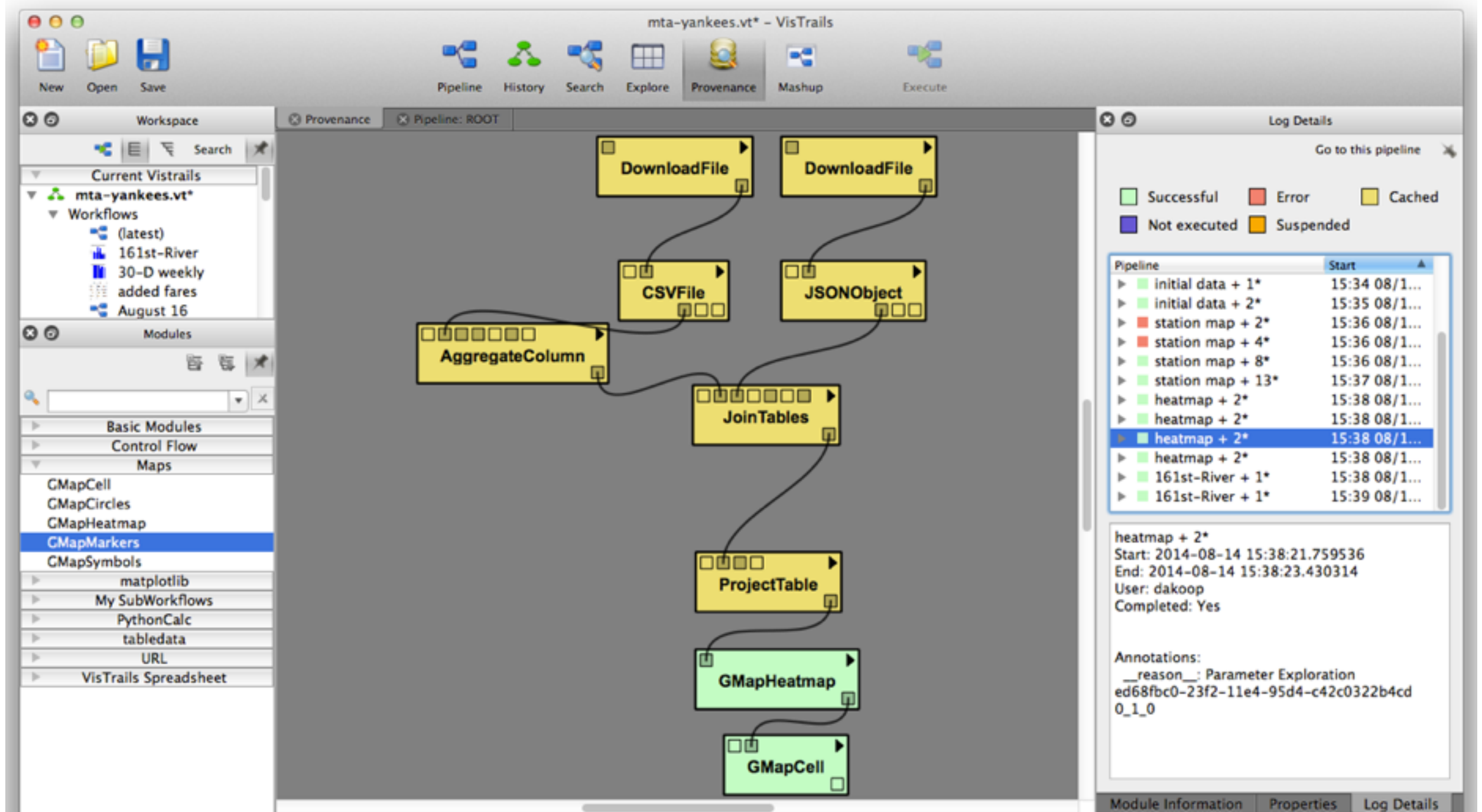
	REMOTE	STATION	FF	T	SEN/DIS	P-D-ARF	U-D-ARF	W-F	J-OBT	RR-TXT	P-D-UNL	30-D-UNL
1	R011	42ND STREET & 8TH AVENUE	00228895	00008471	00000441	00000455	00000134	00003141	00071255			
2	R139	14TH STREET-UNION SQUARE	00224603	00011051	00000427	00000326	00000660	00089167	00109041			
3	R046	42ND STREET & GRAND CENTRAL	00207718	00007908	00000121	00000183	00000300	00040759	00094613			
4	R012	34TH STREET & 8TH AVENUE	00188311	00000640	00000490	00000279	00001822	00035127	00067483			
5	R290	34TH STREET - PENN STATION	00168768	00006155	00000123	00000365	00000831	00030645	00054376			
6	R013	42ND STREET/TIMES SQUARE	00159382	00005945	00000378	00000205	00000690	00018931	00078644			
7	R032	34TH STREET & 6TH AVENUE	00156008	00006276	00000487	00000143	00000712	00018920	00120466			
8	R084	59TH STREET/COLUMBUS CIRCLE	00151262	00009484	00000189	00000207	00000542	00013387	00113966			
9	R029	47-50 STREETS/ROCKEFELLER	00143200	00006402	00000184	00000159	00000723	00037978	00090745			
10	R179	86TH STREET-LEXINGTON AVE	00142409	00010367	00000470	00000339	00000271	00059328	00125250			
11	R023	34TH STREET & 6TH AVENUE	00138012	00005905	00000348	00000112	00000649	00011131	00075040			
12	R029	PARK PLACE	00121614	00004311	00000287	00000991	00000792	00025404	00065162			
13	R047	42ND STREET & GRAND CENTRAL	00100742	00004273	00000181	00000704	00000243	00022888	00068236			
14	R031	34TH STREET & 7TH AVENUE	00091876	00003990	00000232	00000872	00001419	00024284	00018671			
15	R017	LEXINGTON AVENUE	00094615	00004688	00000190	00000833	00000714	00020018	00015066			
16	R175	8TH AVENUE - 14TH STREET	00084313	00003907	00000286	00000144	00000216	00018272	00076681			
17	R057	BAYCLAY'S CENTER	00081804	00004204	00000454	00000186	00000490	00039113	00068119			
18	R118	W/37 4TH ST.-WASH/STON SQ	00081563	00004677	00000351	00000965	00000137	00015628	00074458			



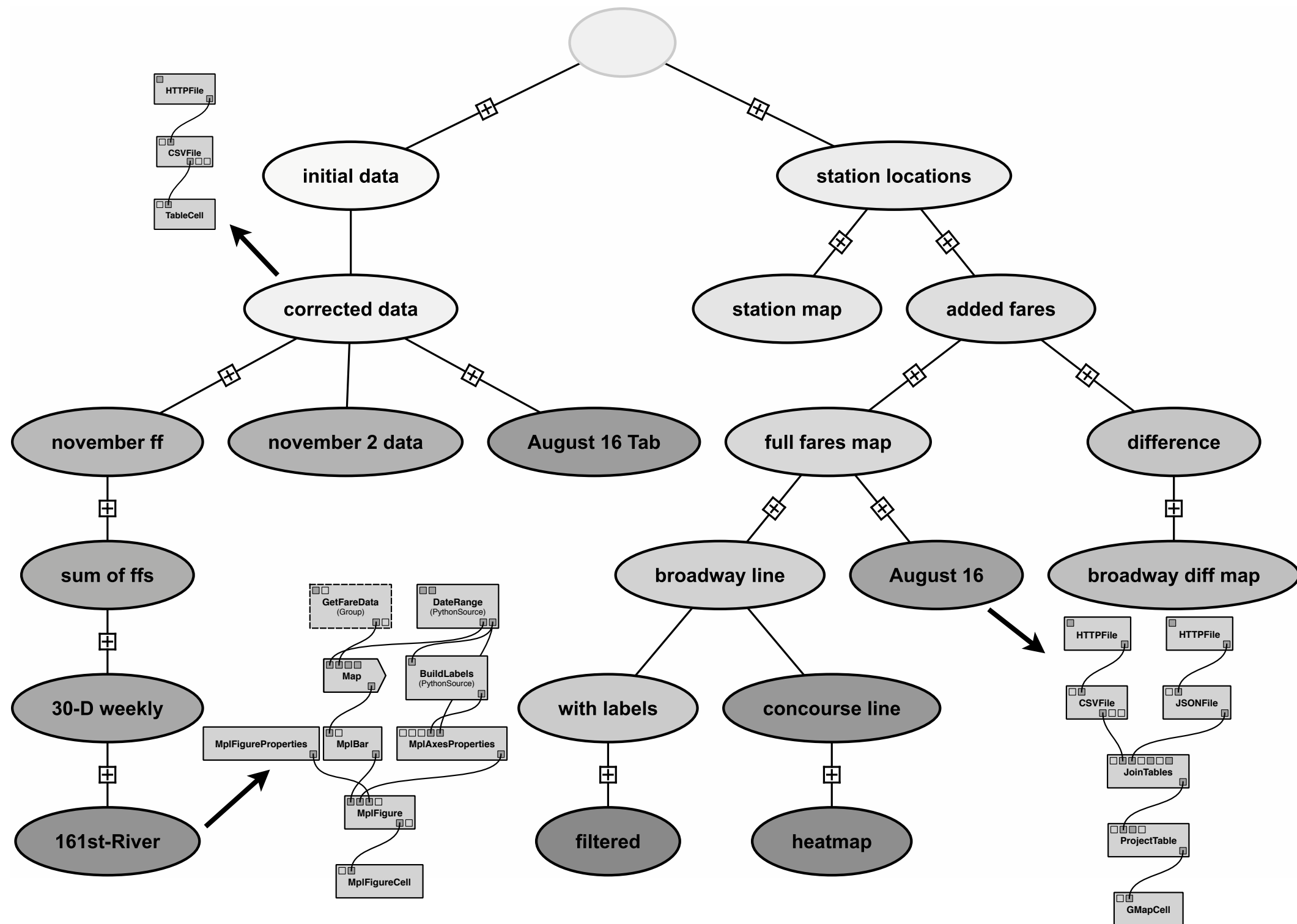
# David Koop



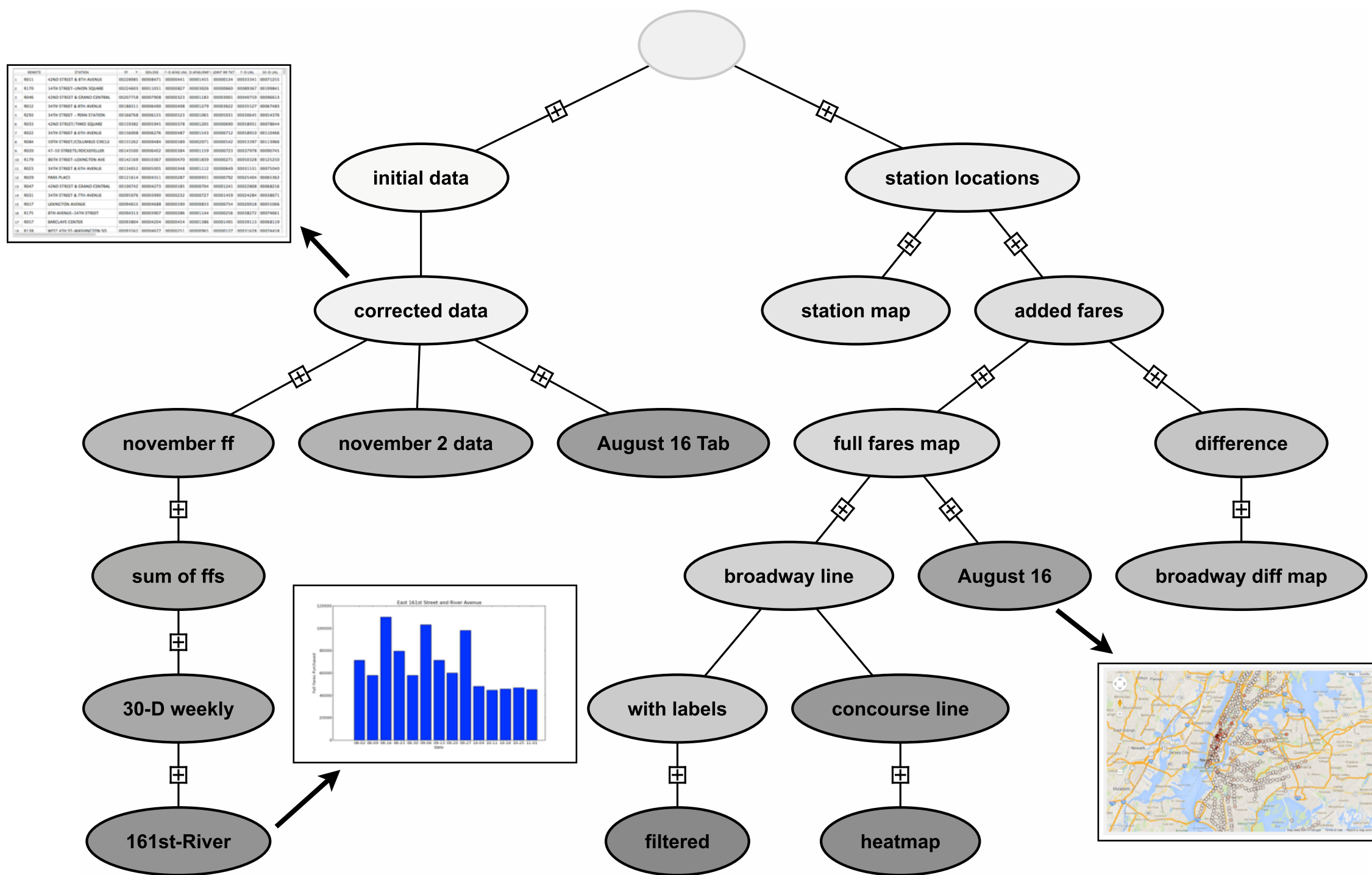
# Execution Provenance



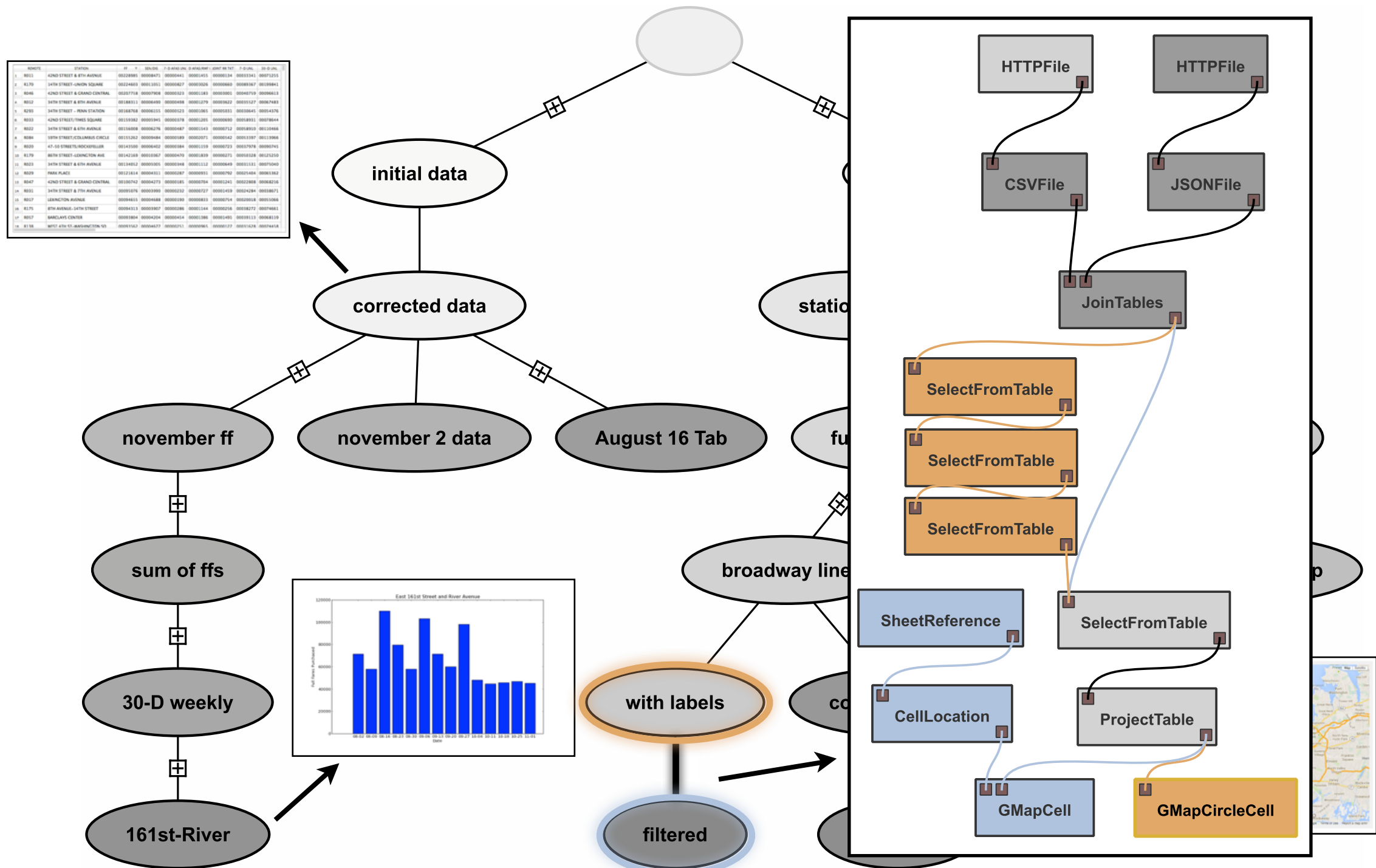
# Workflow Evolution Provenance



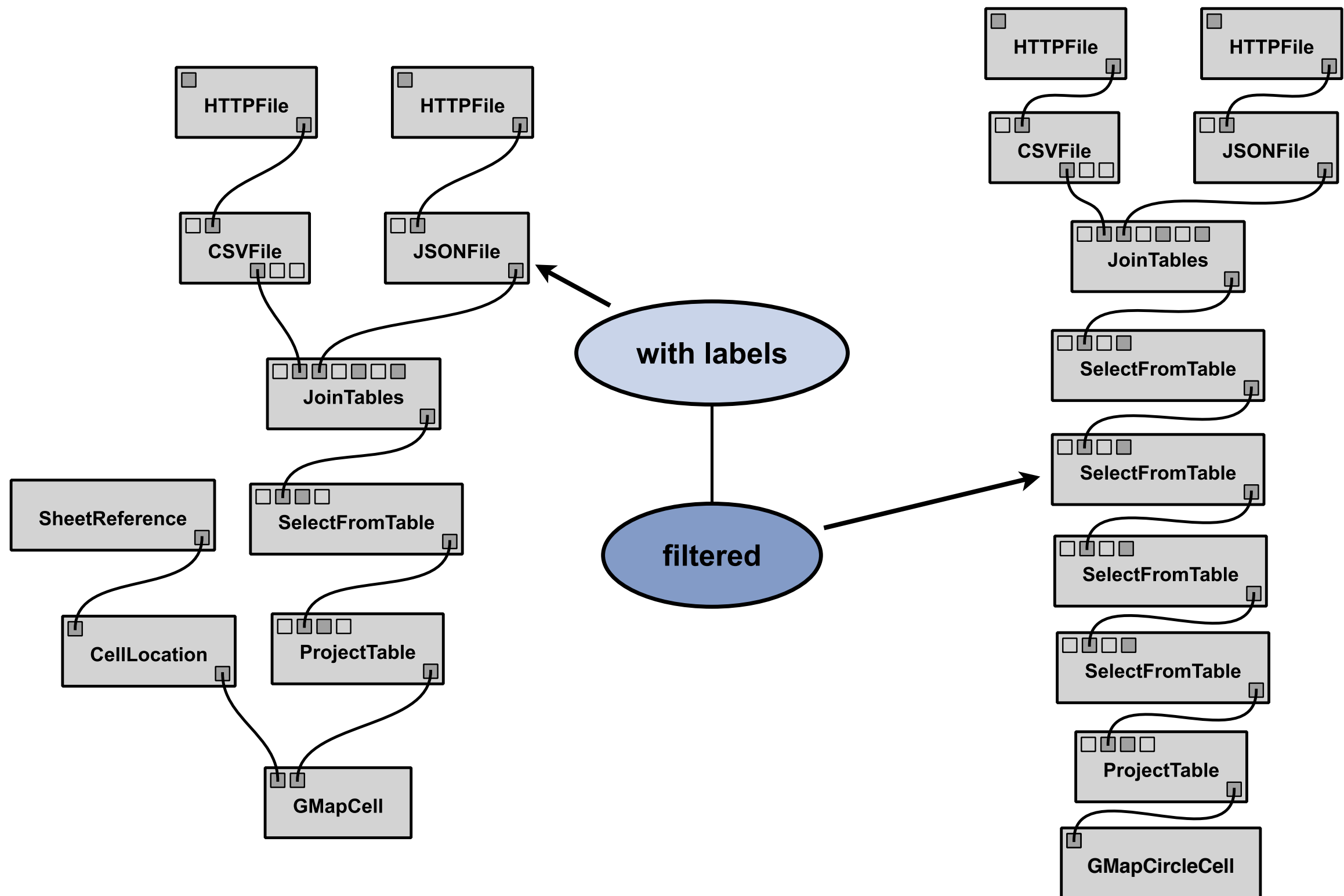
# Workflow Evolution Provenance



# Workflow Evolution Provenance

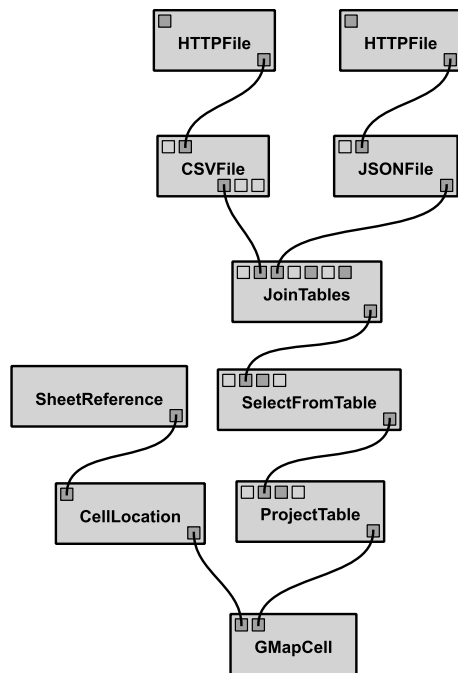
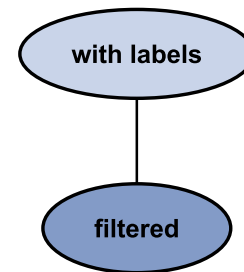


# Workflow Evolution Provenance

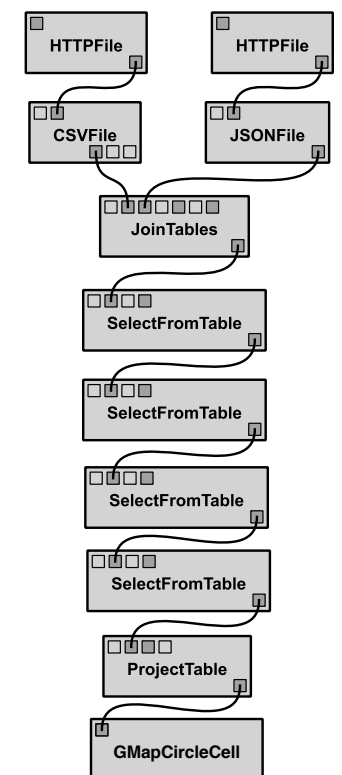




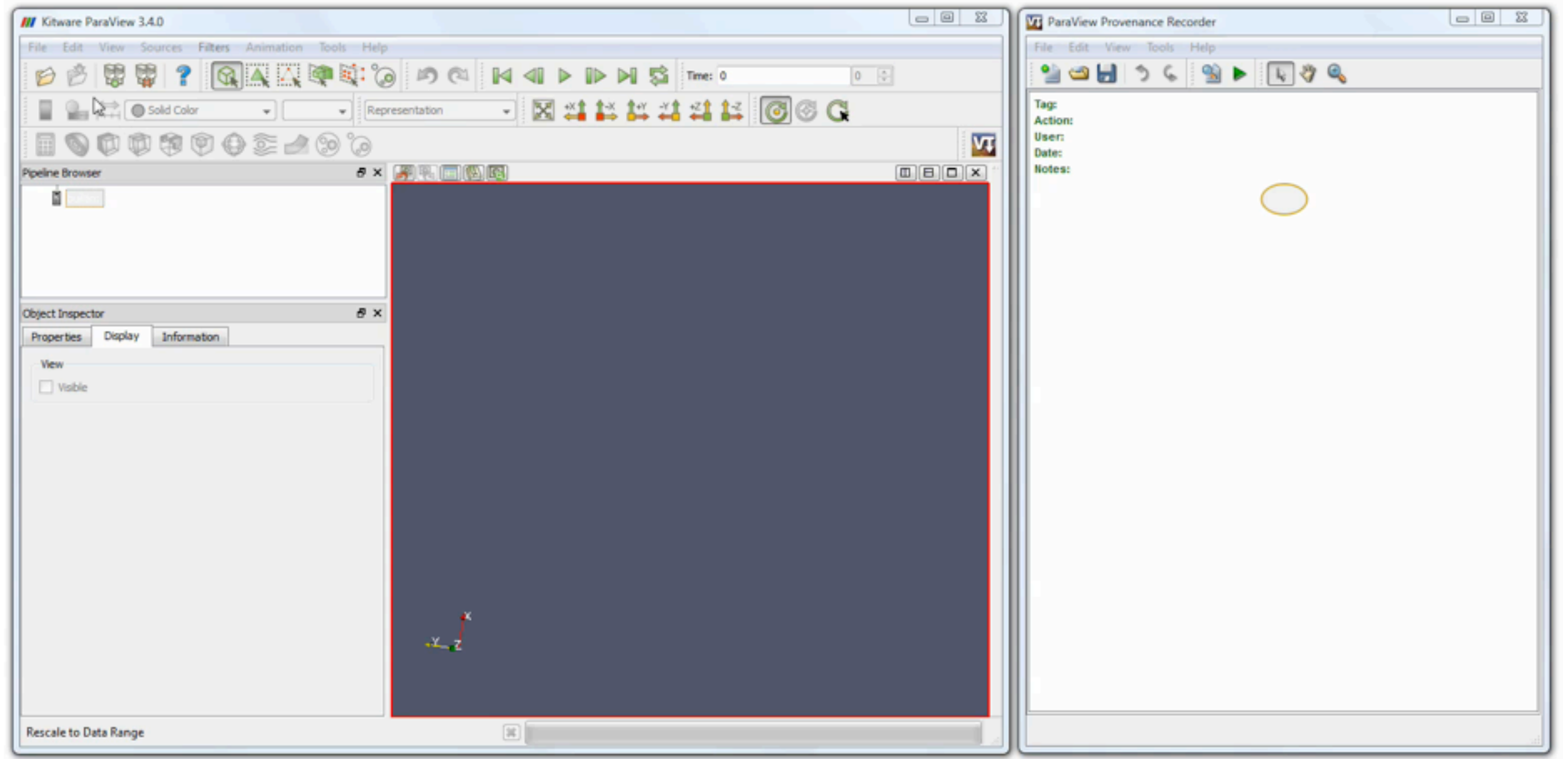
# Workflow Evolution Provenance



delete module "GMapCell"
delete module "CellLocation"
delete module "ProjectTable"
delete module "SelectFromTable"
...
add module "SelectFromTable"
add parameter "float_expr" to "SelectFromTable" with value "latitutde > 40.6"
delete parameter "float_expr" from "SelectFromTable"
add parameter "float_expr" to "SelectFromTable" with value "latitutde > 40.7"
delete parameter "float_expr" from "SelectFromTable"
add parameter "float_expr" to "SelectFromTable" with value "latitutde > 40.8"
...

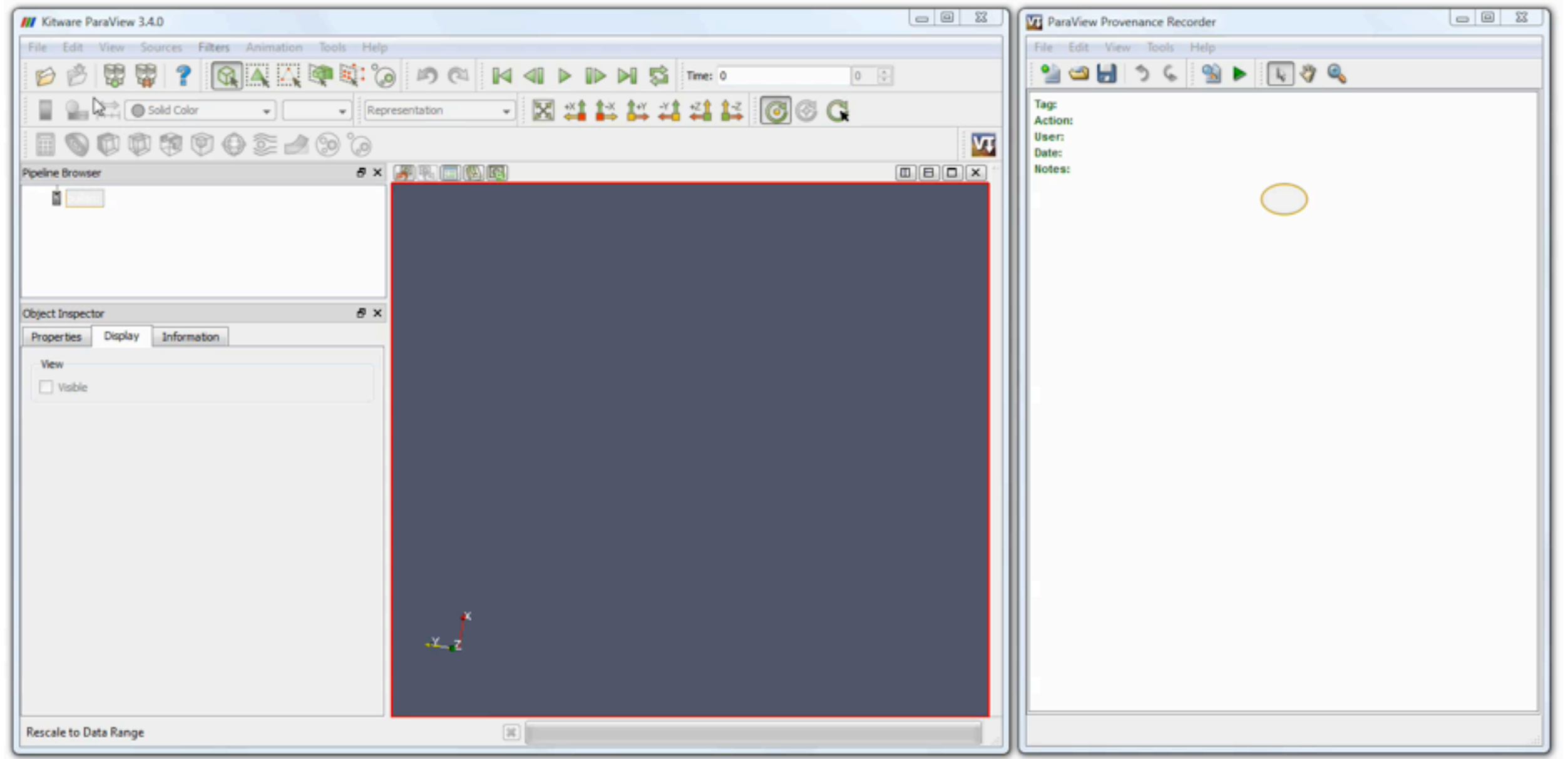


# Evolution Provenance for ParaView



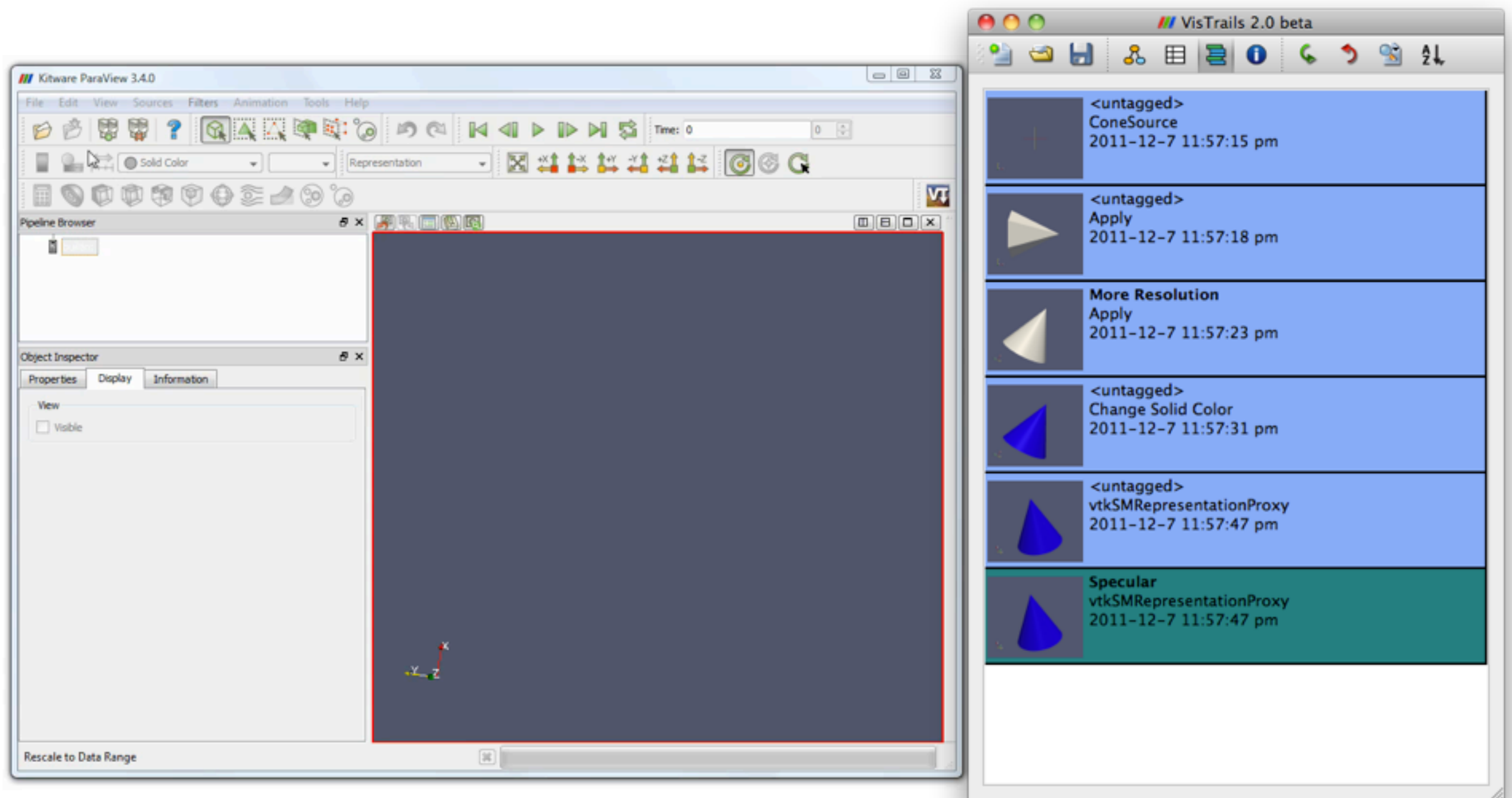
[VisTrails, Inc.]

# Evolution Provenance for ParaView



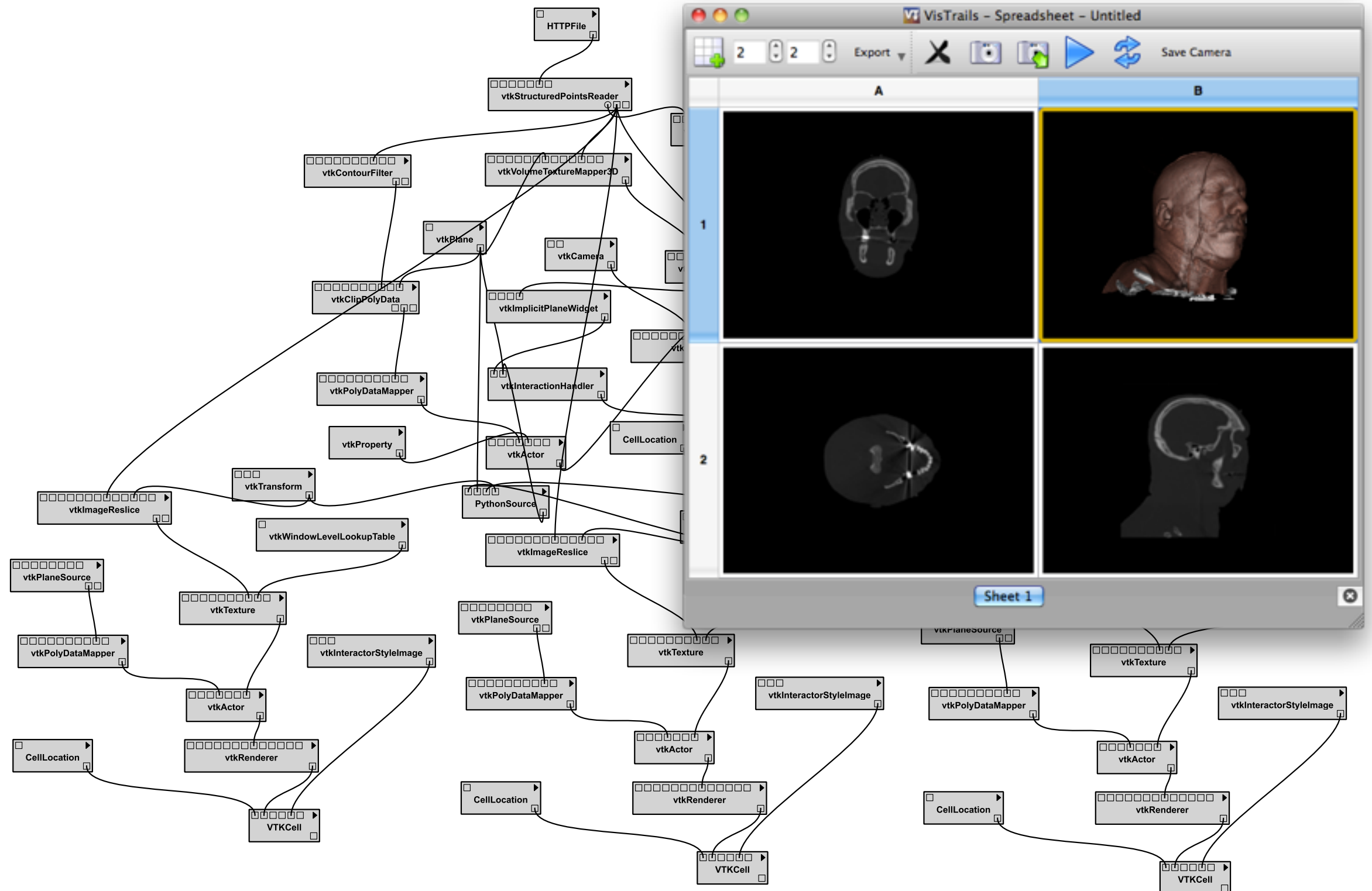
[VisTrails, Inc.]

# Evolution Provenance for ParaView



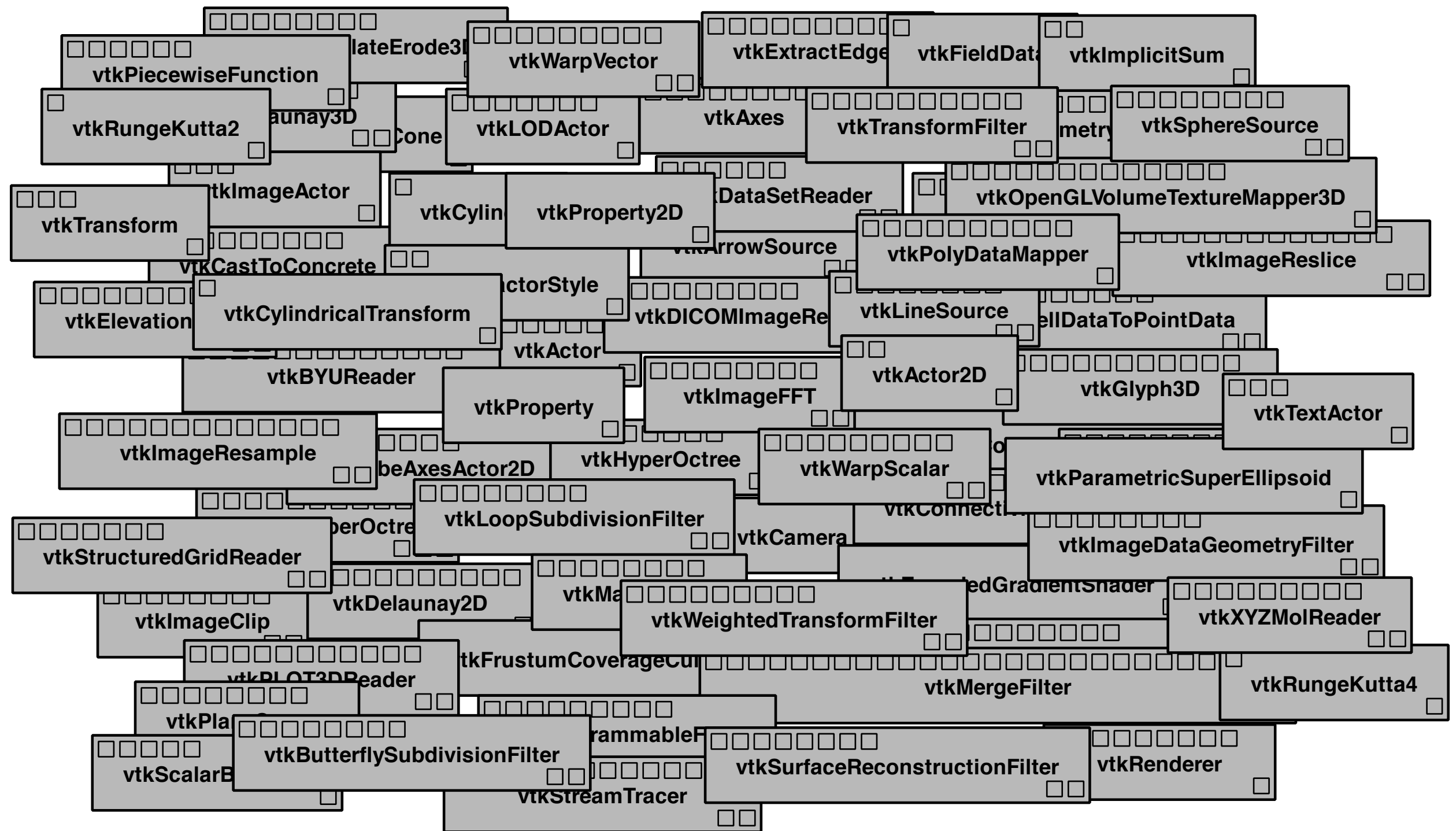
[VisTrails, Inc.]

# Building Visualization Pipelines

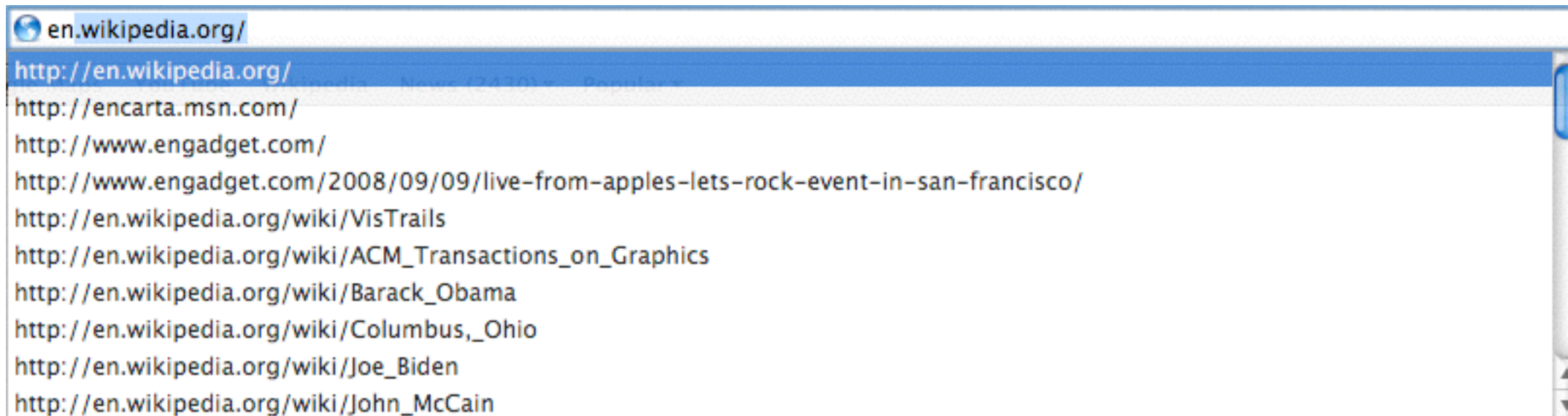




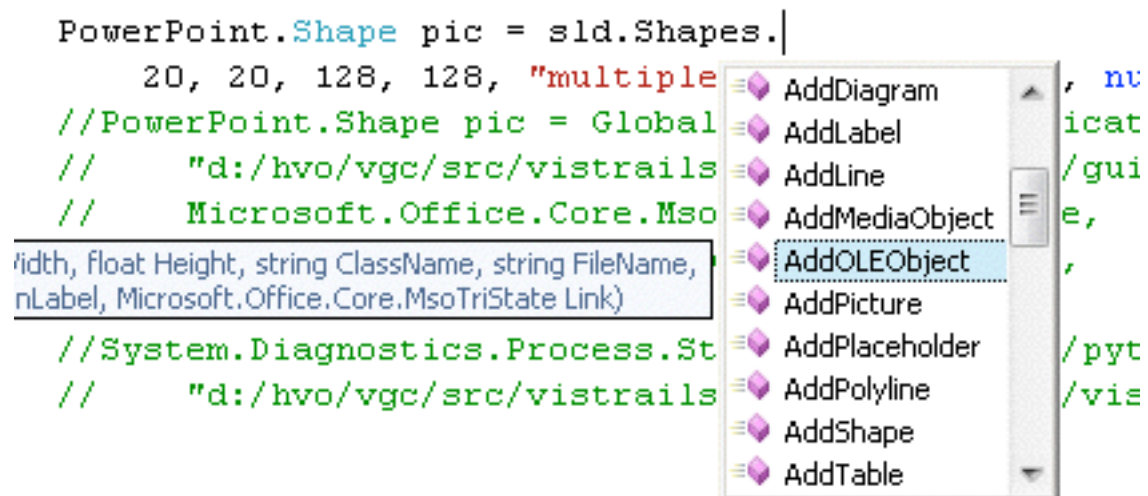
# Building Visualization Pipelines



# Completions



[URL Completion, Safari]

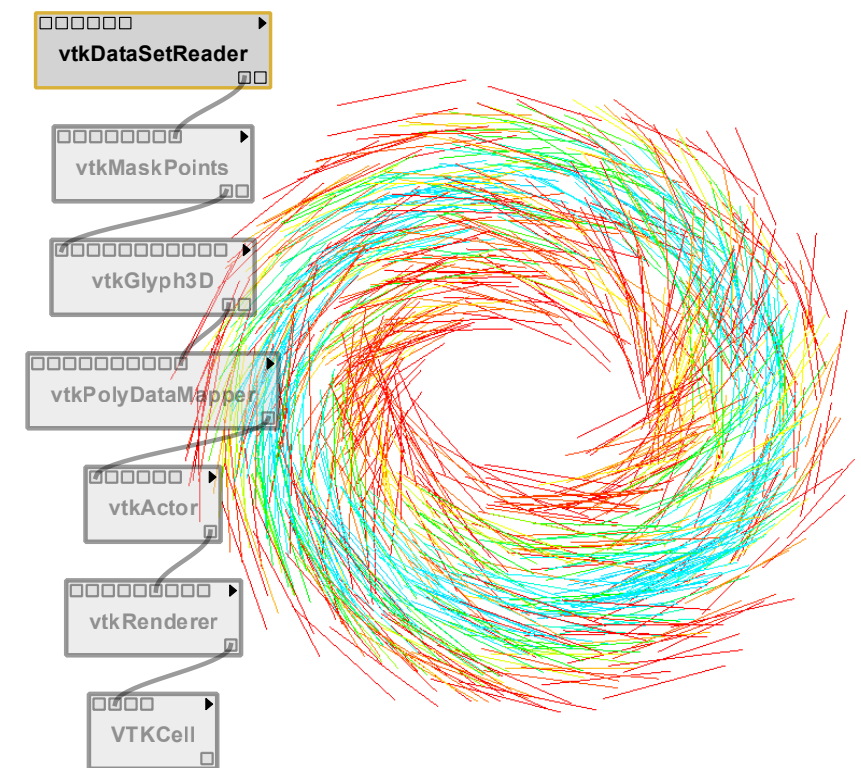
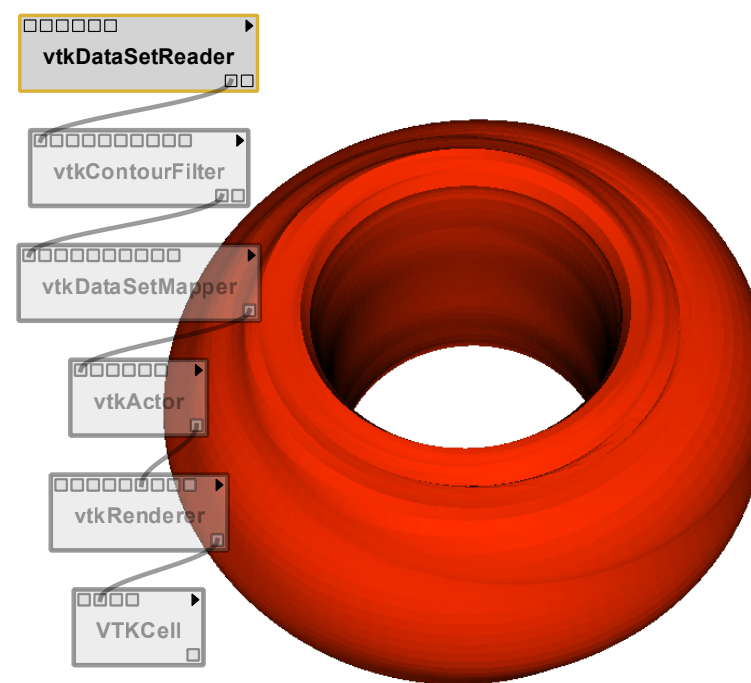
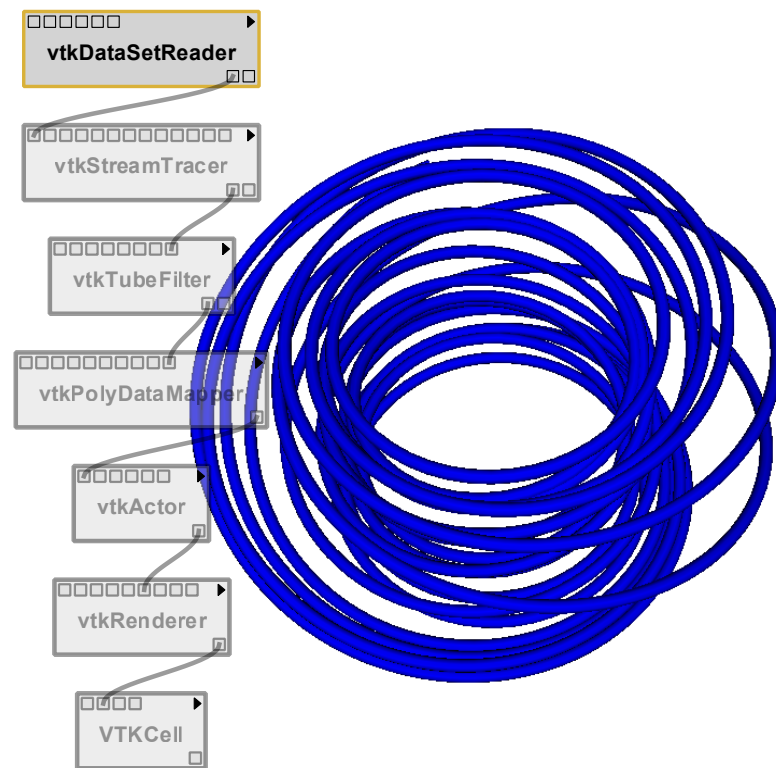


[Code Completion, Intellisense]

visualization	
visualizations for windows media player	1,670,000 results
visualization techniques	954,000 results
visualization tools	2,090,000 results
visualization board	3,380,000 results
visualization api	2,210,000 results
visualization toolkit	368,000 results
visualization technique	756,000 results
visualizations photography	1,830,000 results
visualization meditation	190,000 results
visualizations for media player	1,050,000 results
<a href="#">close</a>	

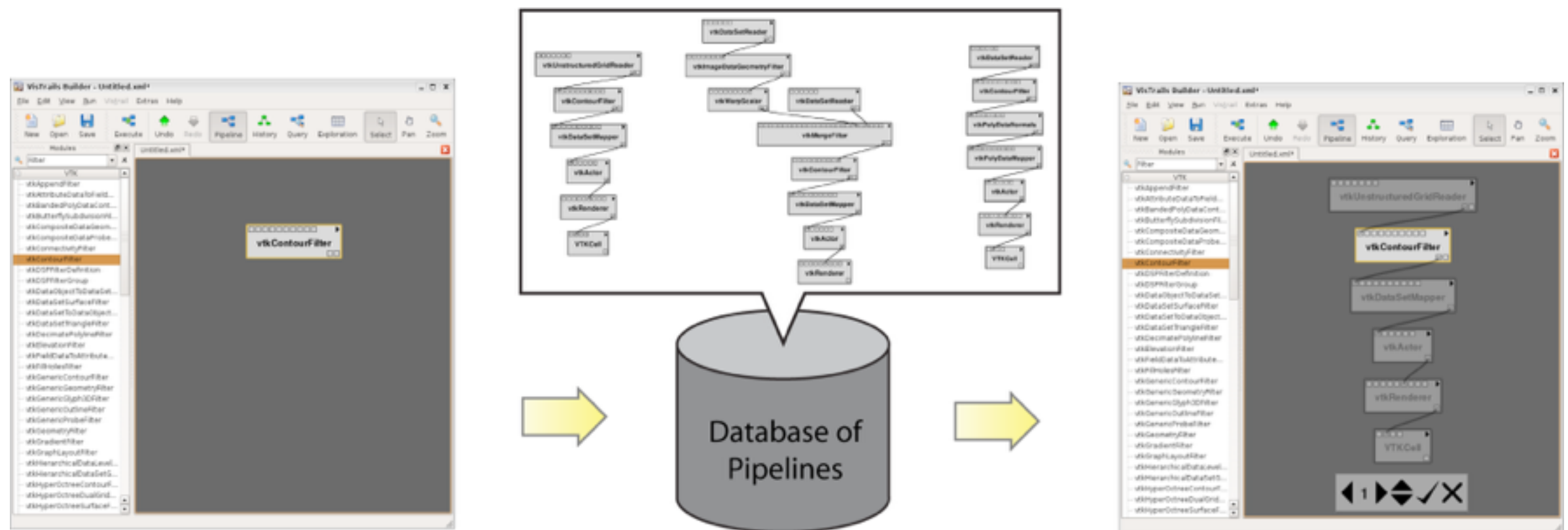
[Web Search Completion, Google]

# Visualization Pipeline Completions



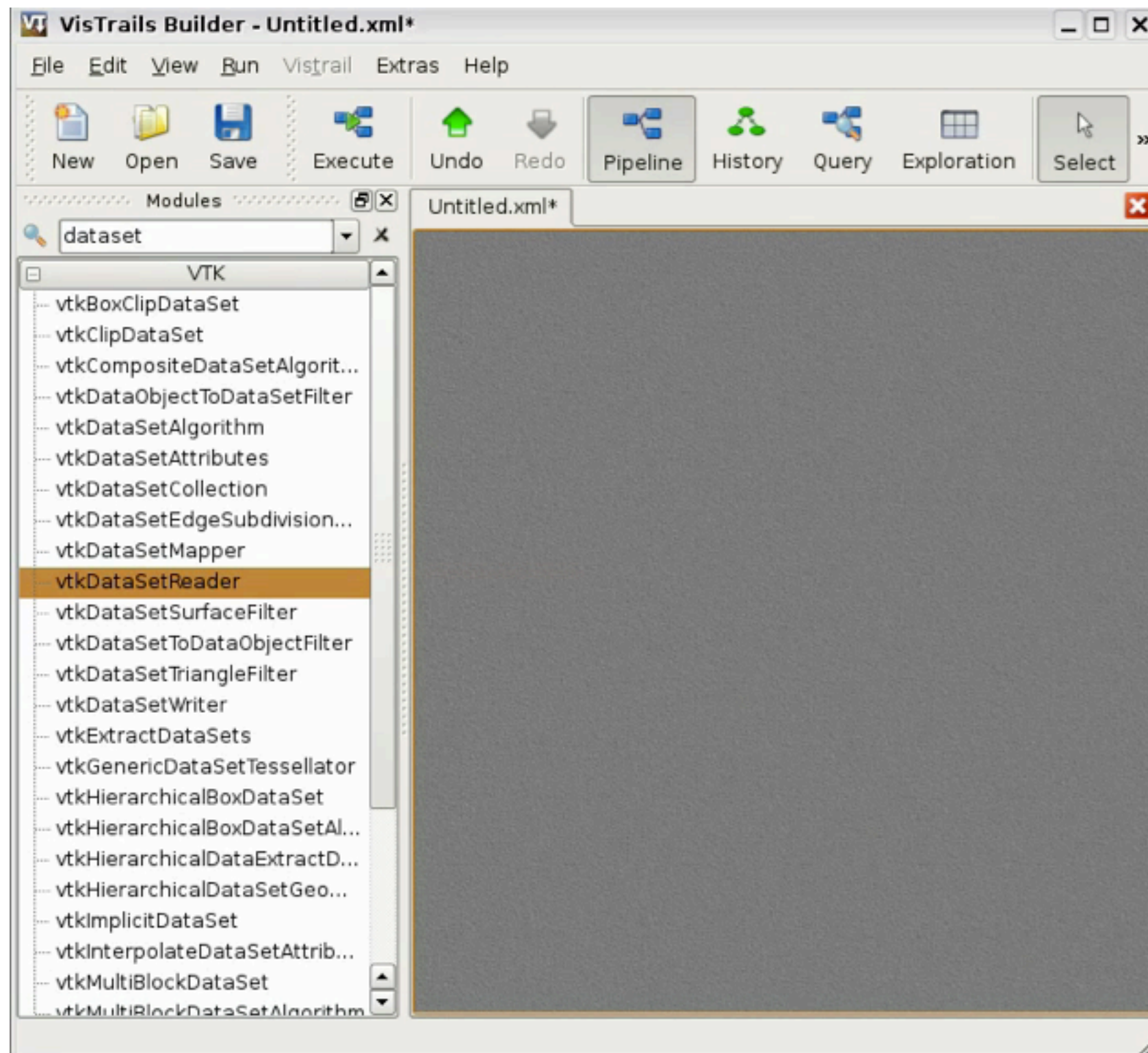
# VisComplete Overview

- Mine provenance collection: Identify graph fragments that co-occur in a collection of workflows (Data-Driven)
- Predict sets of likely workflow additions to a given partial workflow



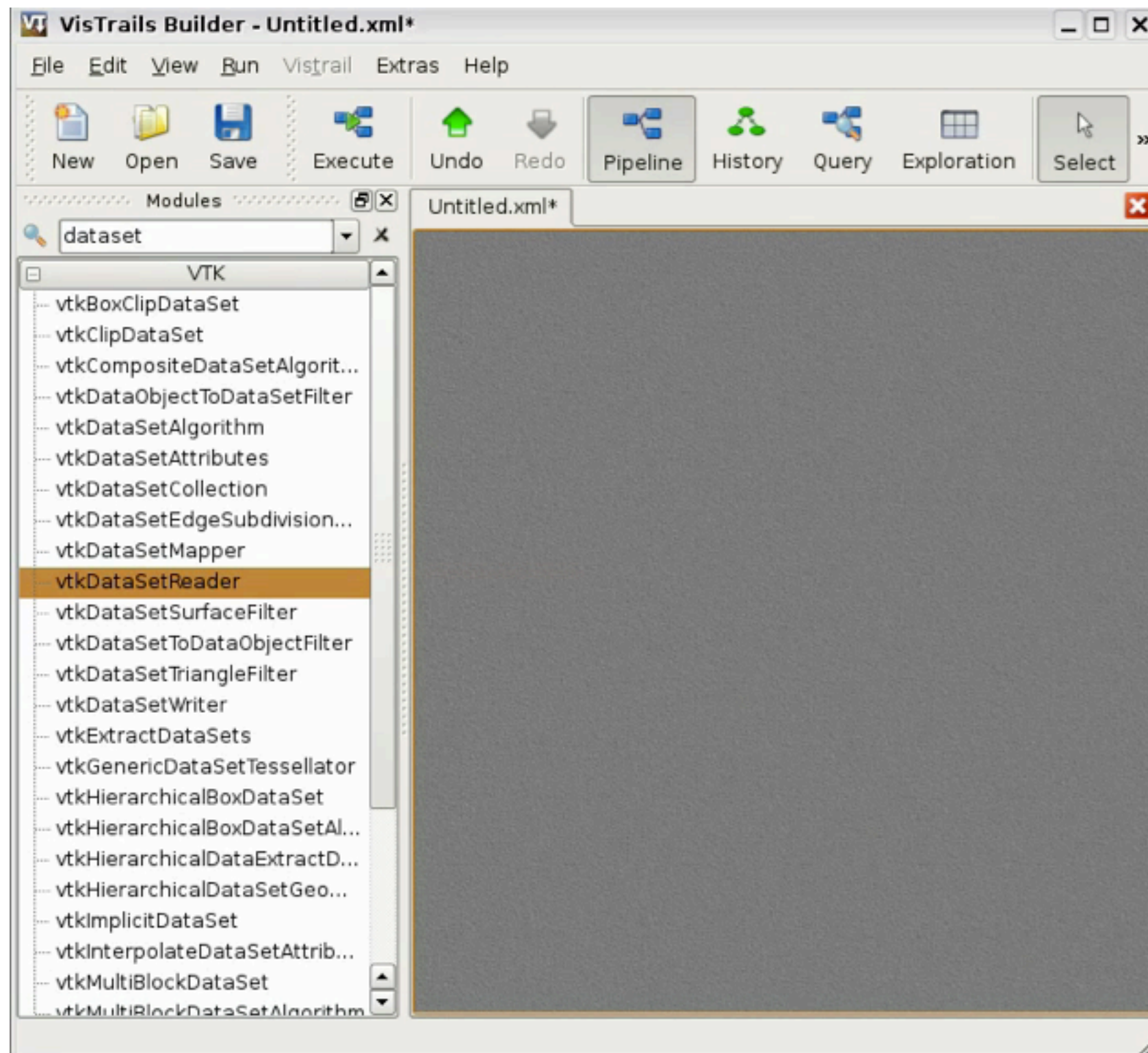


# Suggestion Interface

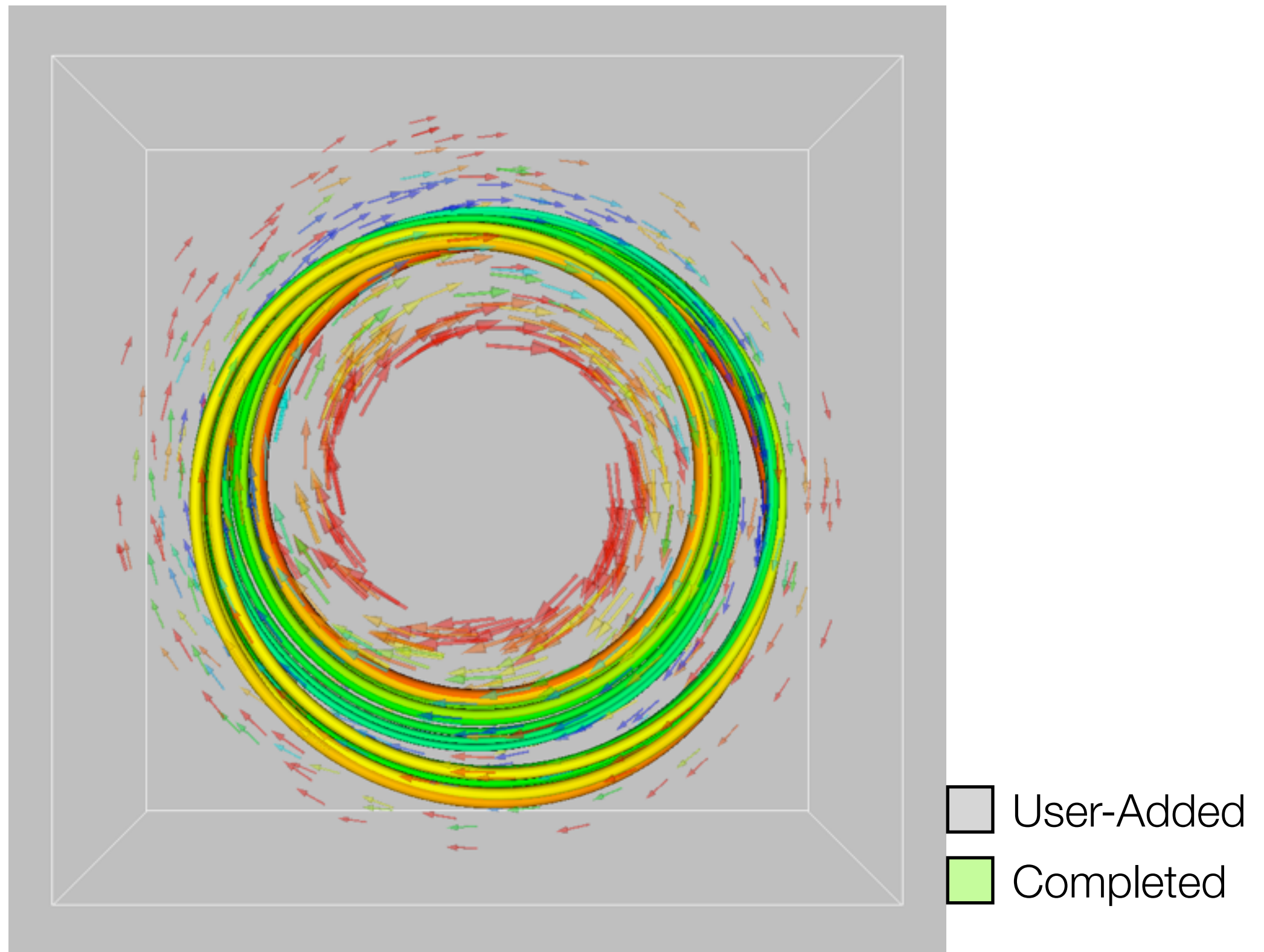




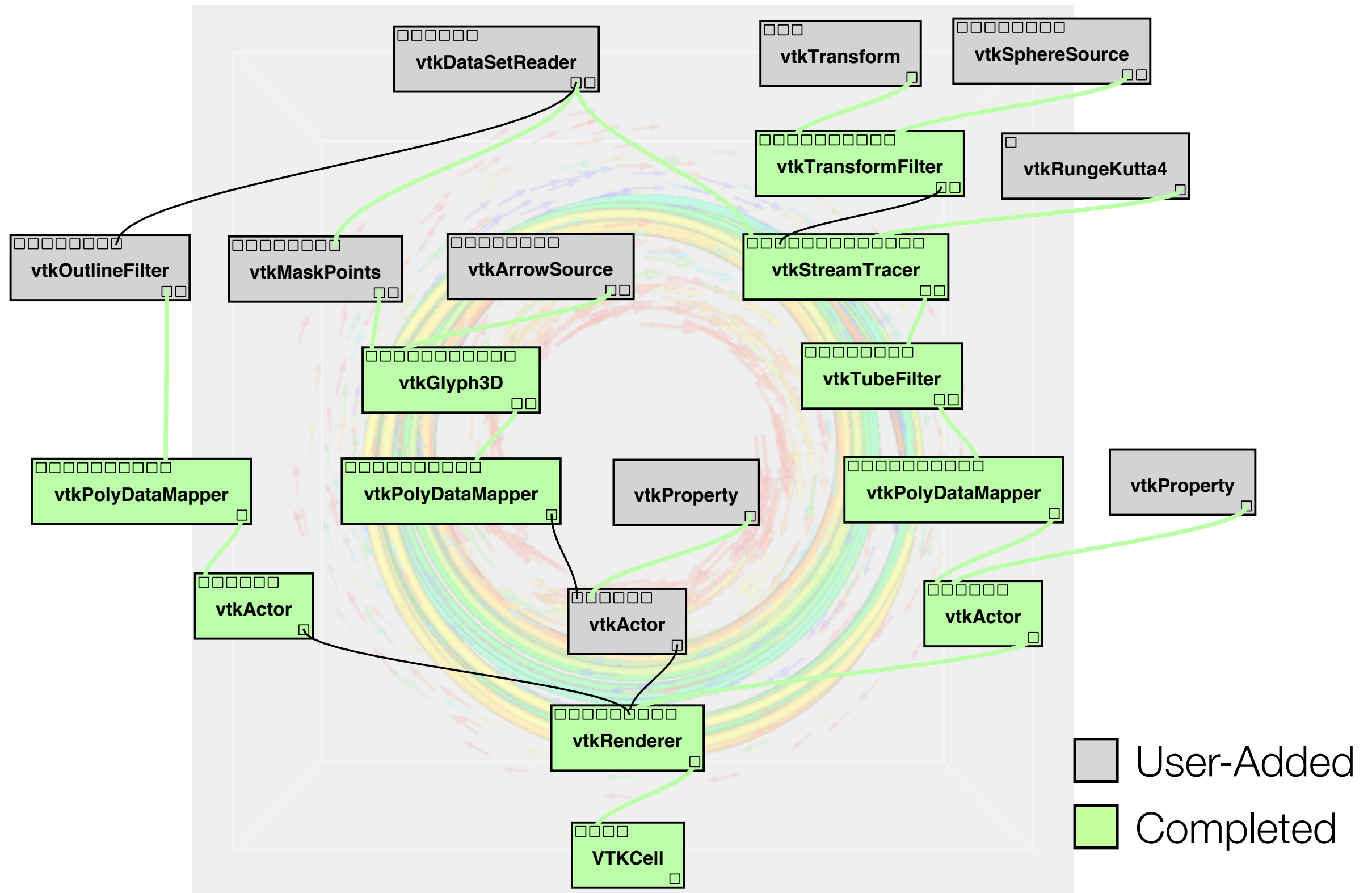
# Suggestion Interface



# VisComplete Results

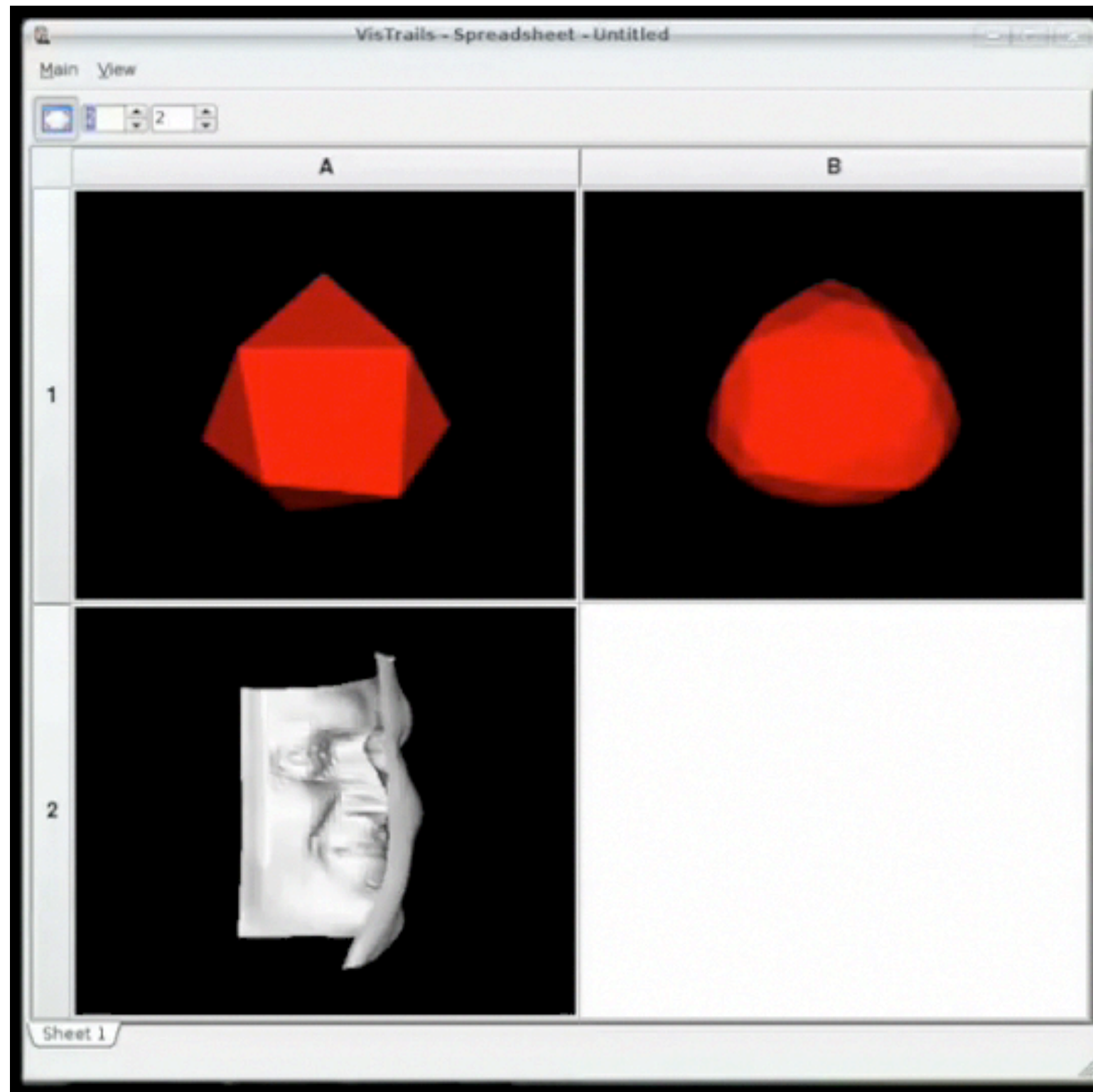


# VisComplete Results

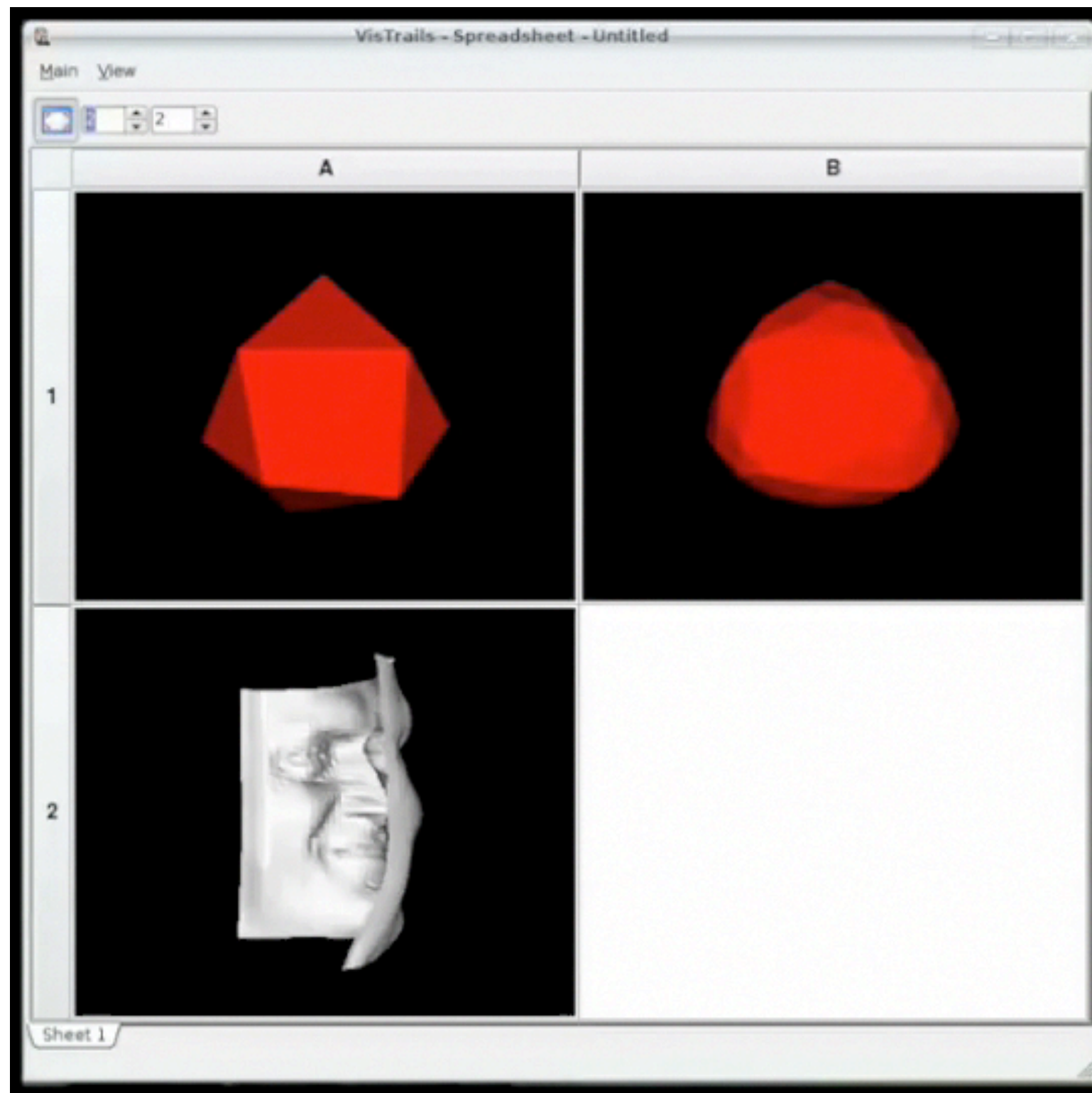




# Visualization by Analogy

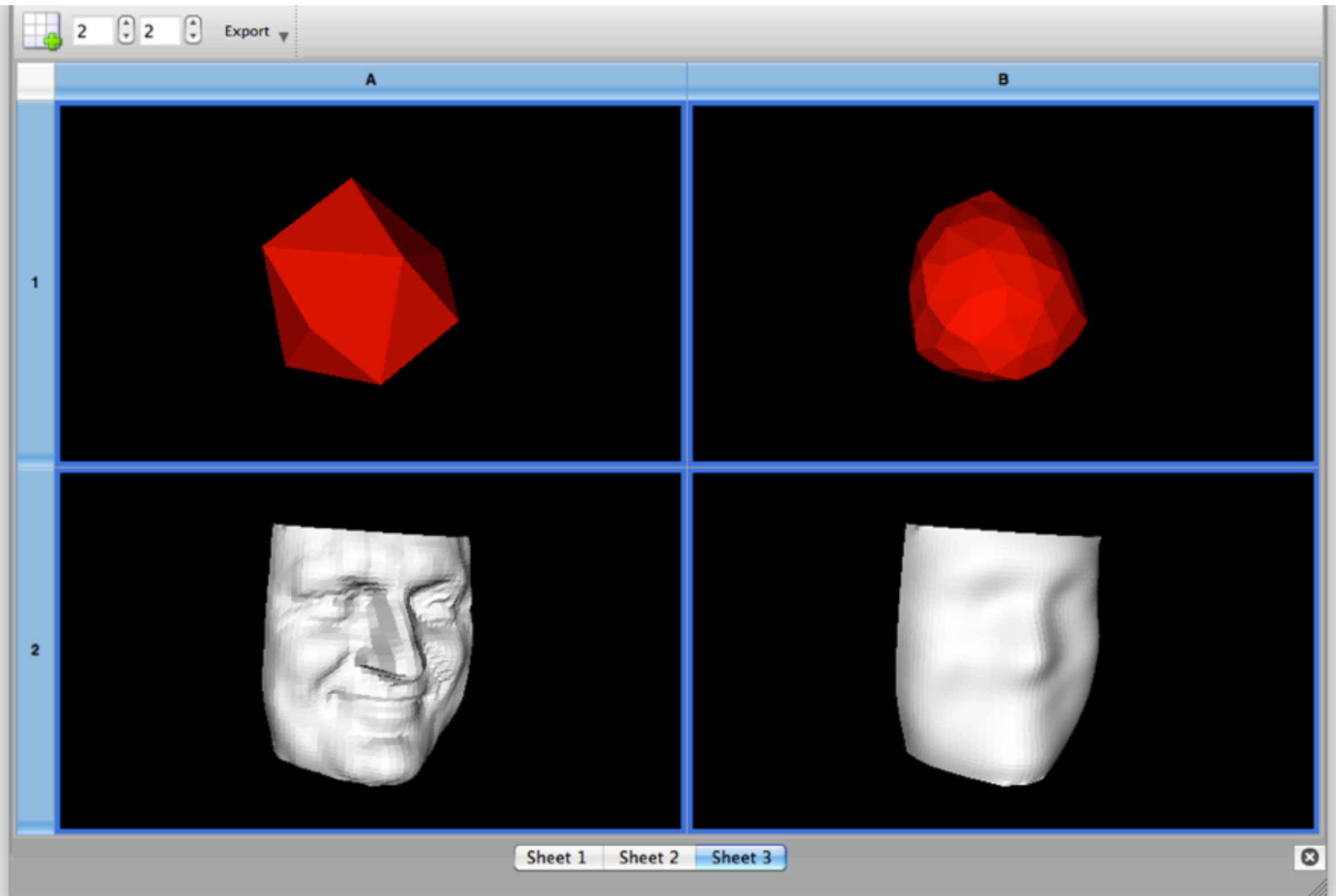


# Visualization by Analogy

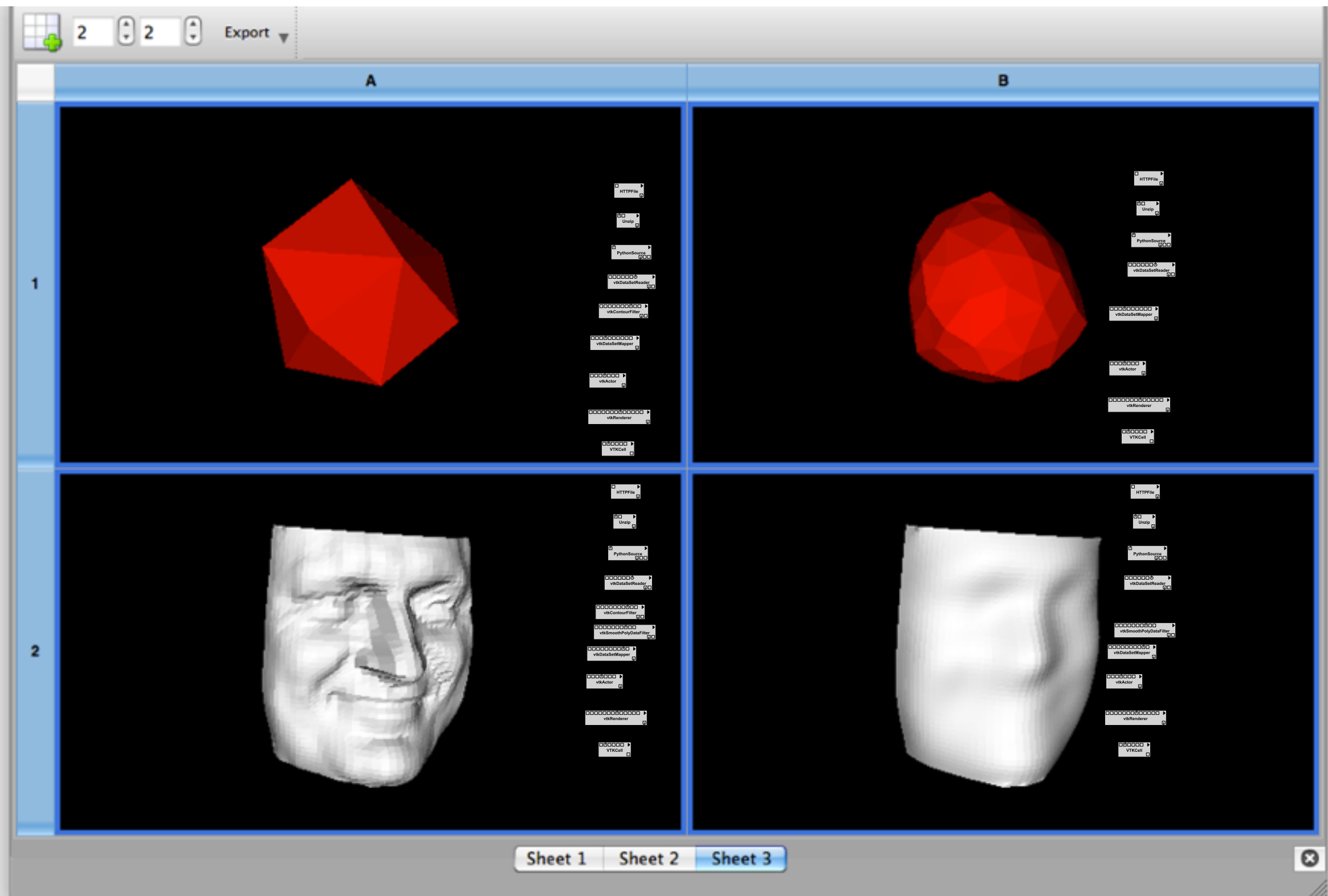




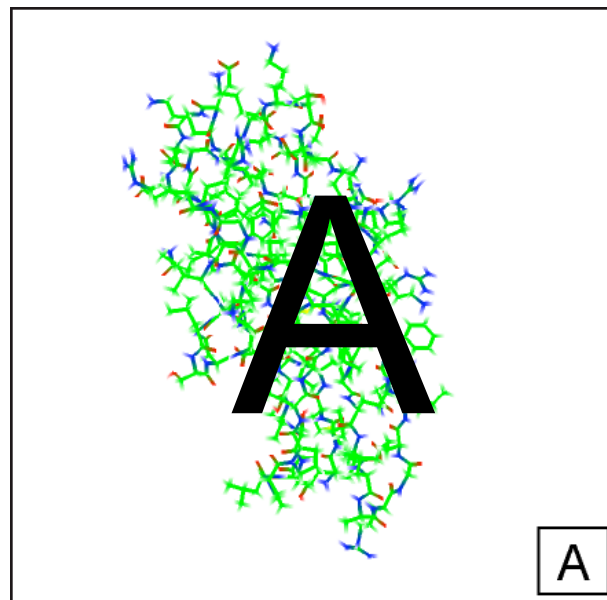
# Visualization by Analogy



# Visualization by Analogy

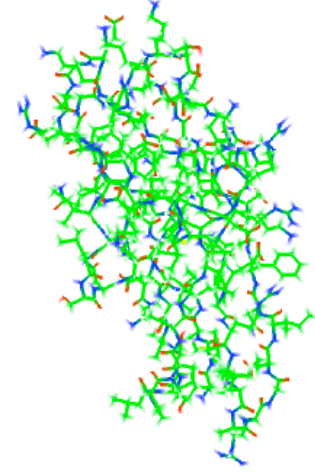


# Generating Visualizations by Analogy



is to

PDB Report

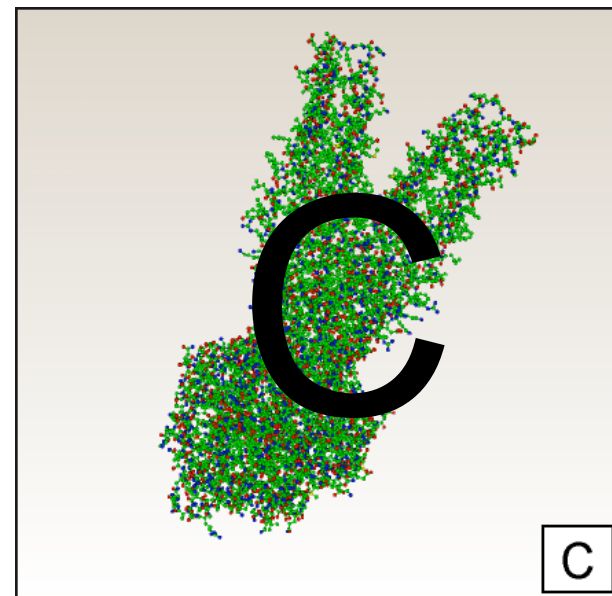


**B**

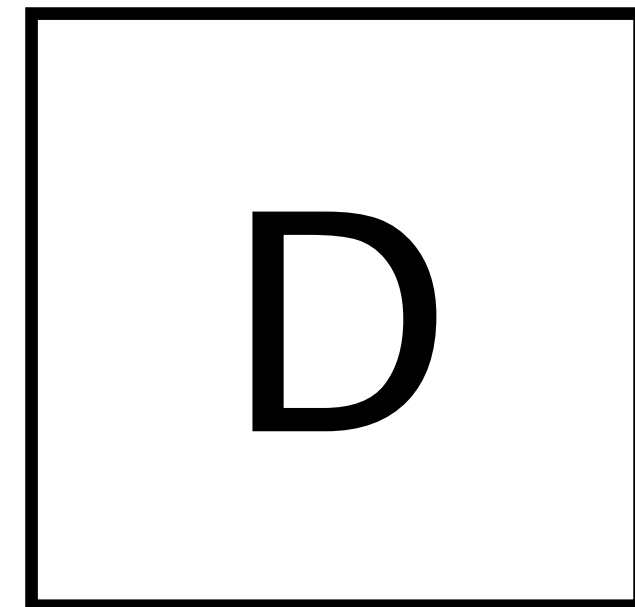
<b>Protein Title</b>	NEURAL CELL ADHESION MOLECULE, MODULE 2, NMR, 20 STRUCTURES
<b>Authors</b>	P.H.JENSEN, V.SOROKA, N.K.THOMSEN, V.BEREZIN, E. BOCK, F.M.POULSEN
<b>Atom Count</b>	C: 9560 H: 15440 N: 2580 O: 2680 S: 60
<b>Links</b>	<a href="#">PDB Entry</a>

B

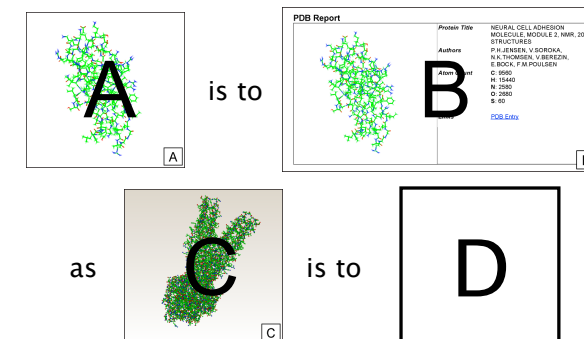
as



is to

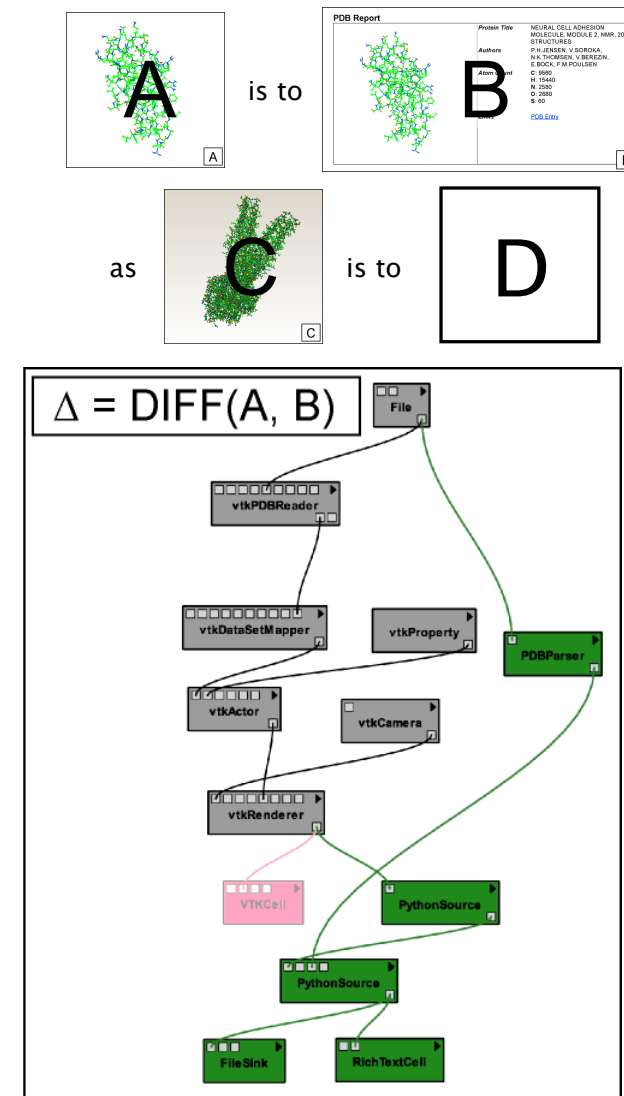


# Generating Visualizations by Analogy



# Generating Visualizations by Analogy

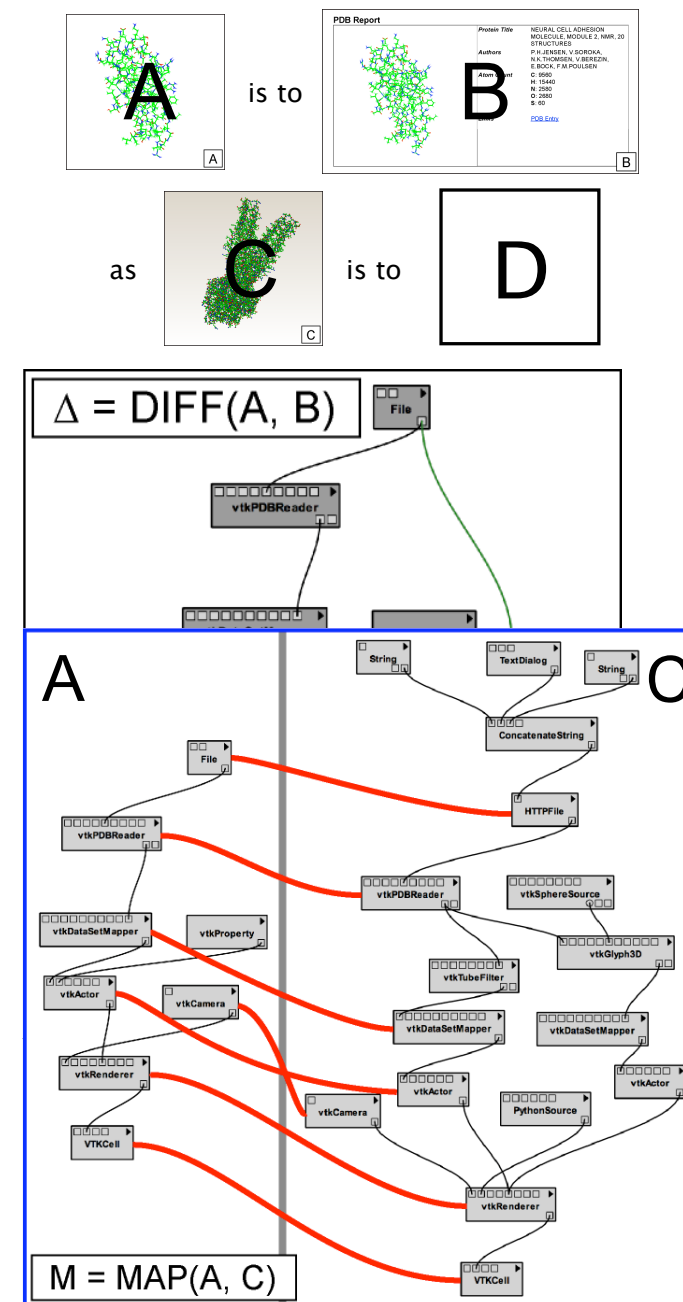
- Compute difference  $\Delta(A,B)$  from provenance
  - $D = \Delta(A,B) \circ C$  is often not a valid workflow





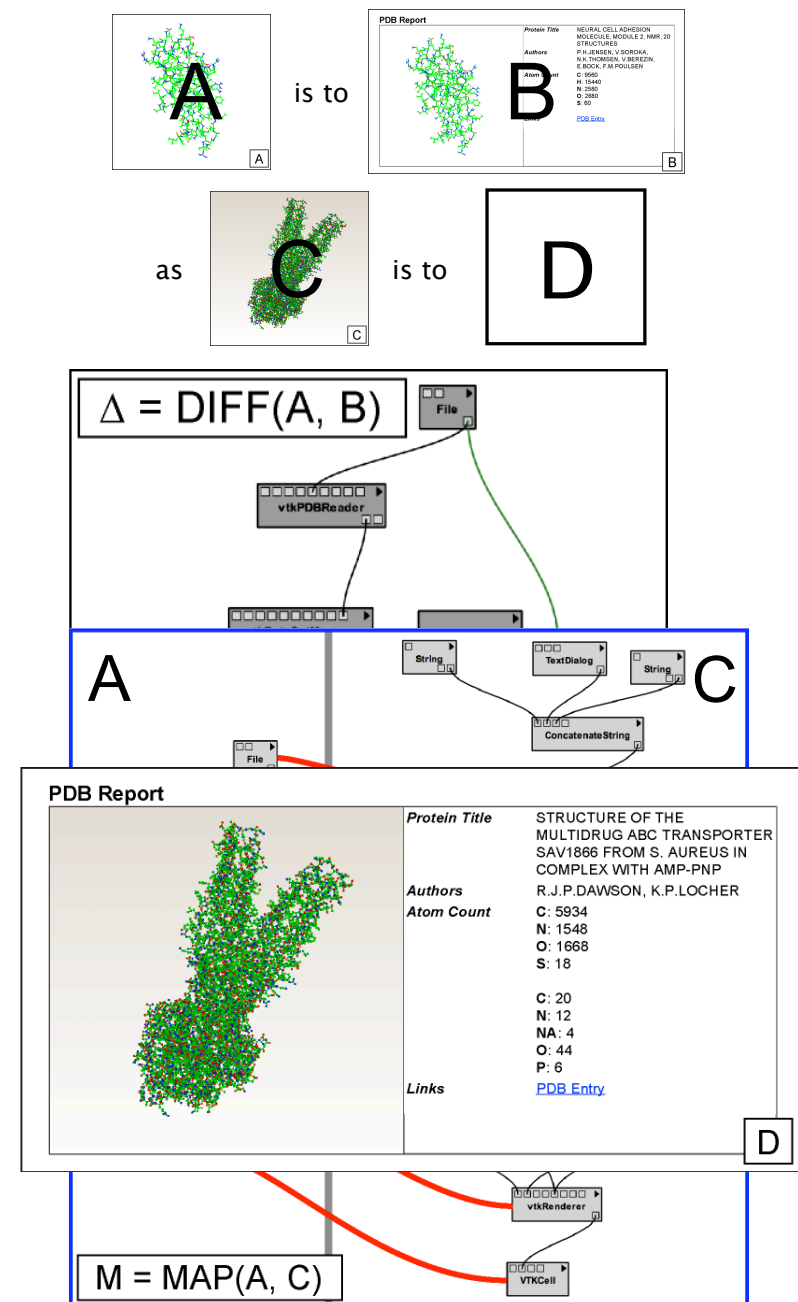
# Generating Visualizations by Analogy

- Compute difference  $\Delta(A,B)$  from provenance
  - $D = \Delta(A,B) \circ C$  is often not a valid workflow
- Find map between A & C:  $\text{map}(A,C)$

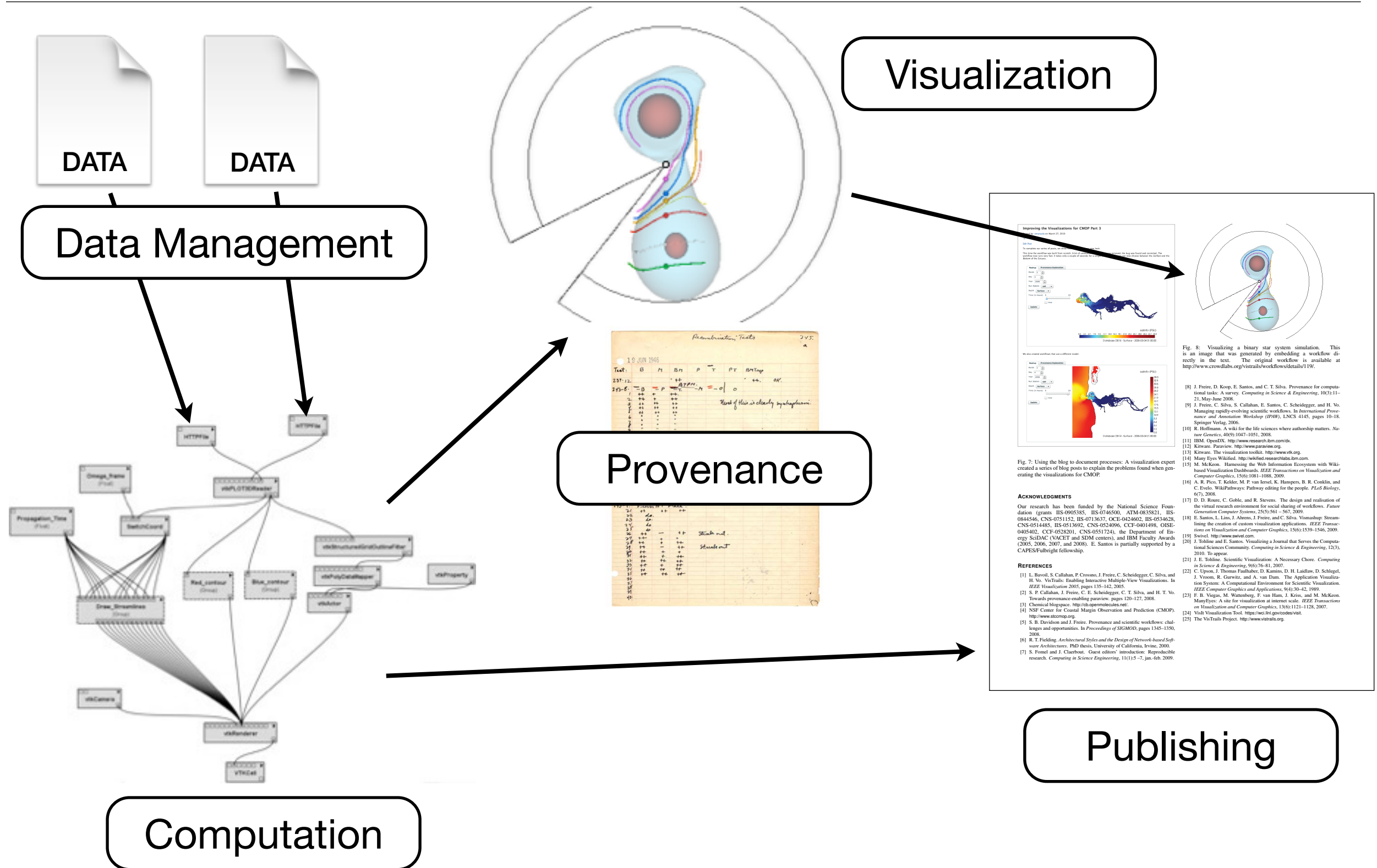


# Generating Visualizations by Analogy

- Compute difference  $\Delta(A,B)$  from provenance
  - $D = \Delta(A,B) \circ C$  is often not a valid workflow
- Find map between A & C:  $\text{map}(A,C)$
- Compute mapped difference
 
$$\Delta_{AC}(A,B) = \text{map}(A,C) \Delta(A,B)$$
  - $D = \Delta_{AC}(A,B) \circ C$



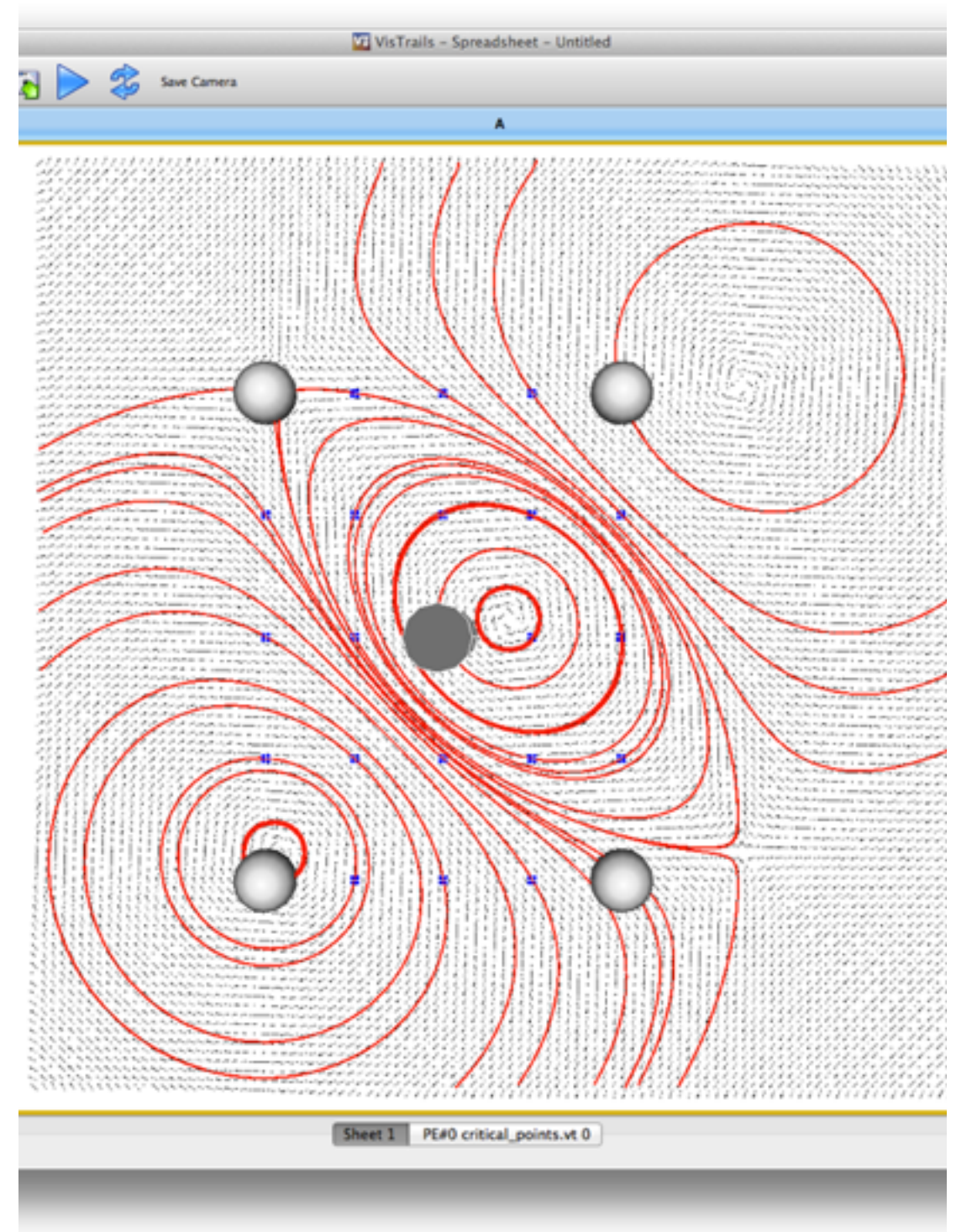
# Vision: Provenance-Rich Science





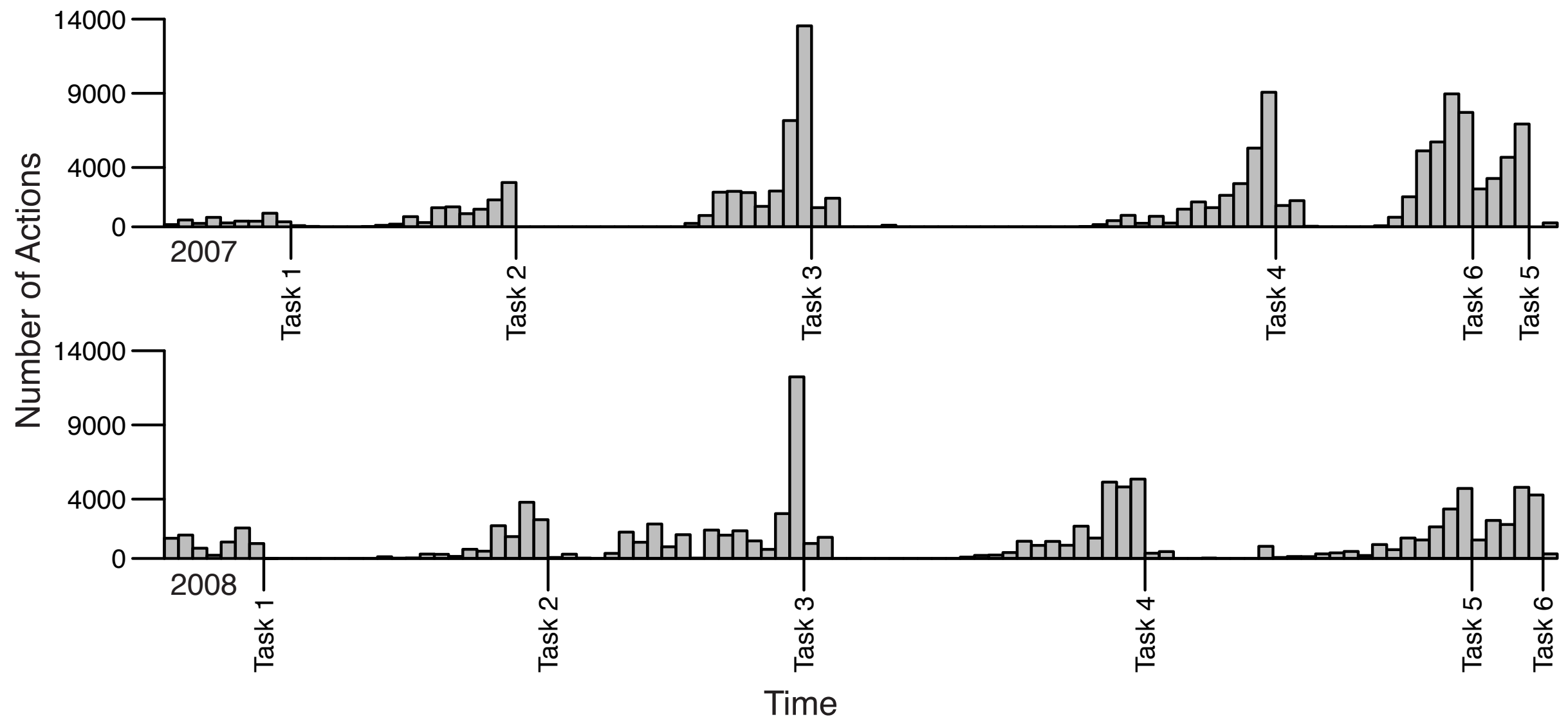
# Provenance in Teaching

- "Using VisTrails and Provenance for Teaching Scientific Visualization" [Silva et al., 2010]
- Same features that scientists use for exploratory tasks can also benefit students
  - Exploration: see all pipelines not just a "final" one
  - Comparison: see different pipelines and what changes exist
  - Assessment: see how a solution was developed



# Provenance Analysis of Projects

## Activity Histograms by Date

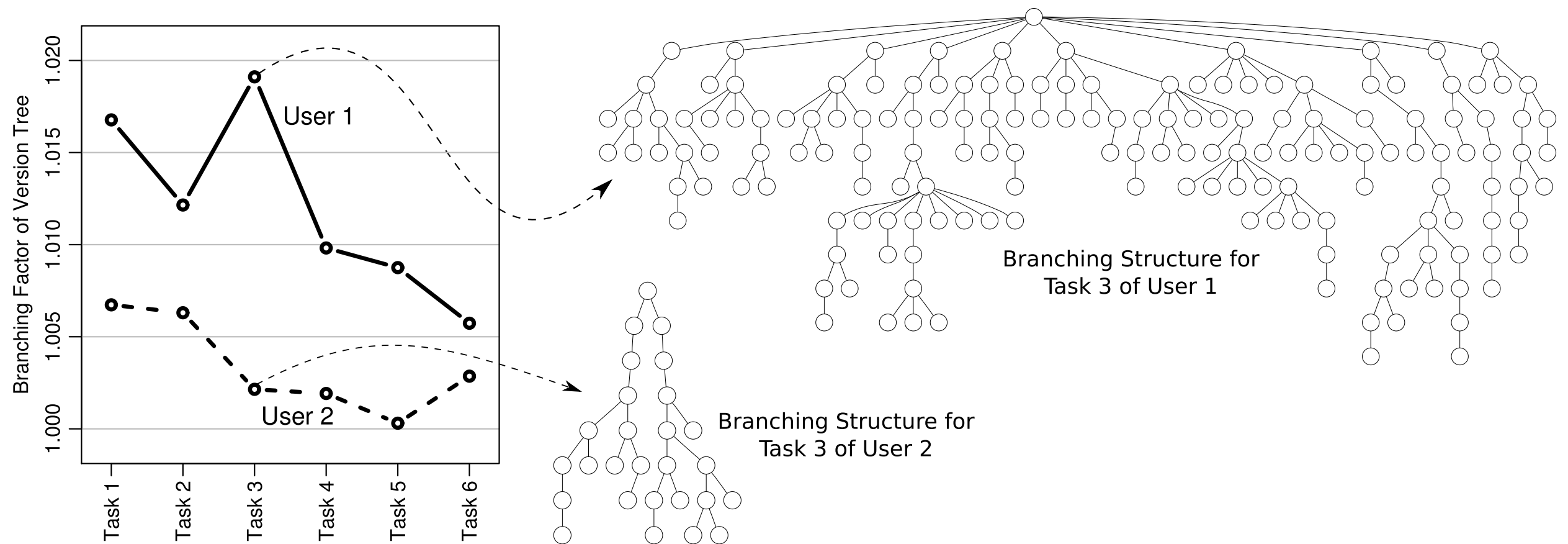


[Lins et al., 2008]



# Provenance Analysis of Projects

## Comparing Paths to Solutions for Two Students

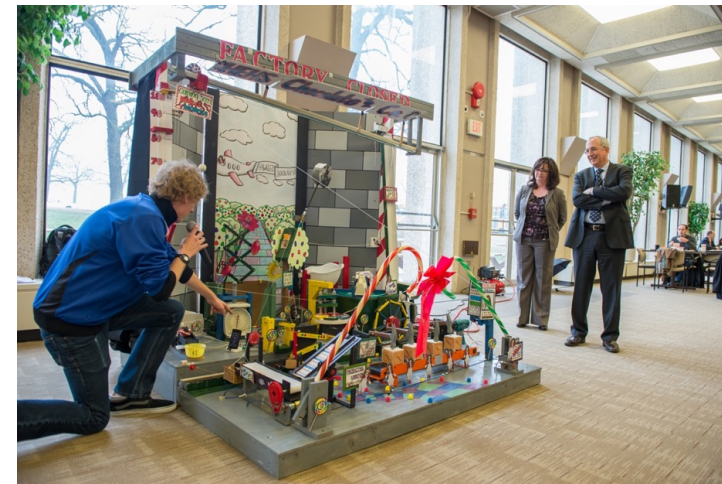


[Lins et al., 2008]

# Conclusion: Uses of Provenance



Trust



Reproducibility



Reuse



Exploration

# Conclusion: Take-Home Points

---

- Provenance is important
- Consider what needs to be stored, what would be nice to store, and what doesn't need to be stored
- Abstraction helps
- PROV model organizes and connects a variety of provenance
- Provenance isn't just for keeping a paper trail

# Conclusion: References

---

- Bose and Frew: "Lineage Retrieval for Scientific Data Processing: A Survey"
- Simmhan et al.: "A Survey of Data Provenance in E-Science"
- Tan, "Provenance in Databases: Past, Current, and Future"
- Freire et al.: "Provenance for Computational Tasks: A Survey"
- PROV Model Primer: <http://www.w3.org/TR/prov-primer/>



# Questions

